A Systems-Level Approach to Understand The Seasonal Factors Of Early Development With Clinical and Pharmacological Applications

Mary Regina Boland

Submitted in Partial Fulfillment of the

Requirements For the Degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

A Systems-Level Approach to Understand the Seasonal Factors of Early Development with Clinical and Pharmacological Applications

Mary Regina Boland

Major developmental defects occur in 100,000 to 200,000 children born each year in the United States of America. 97% of these defects are from unidentified causes. Many fetal outcomes (e.g., developmental defects), result from interactions between genetic and environmental factors. The lifetime effects from prenatal exposures with low impact (e.g., air pollution) are often understudied. Even when these exposures are studied, the focus is often placed on immediate effects of the exposure (e.g., fetal anomalies, miscarriage rates) leaving lifetime effects largely unexplored. This makes prolonged (or lifetime) effects of low-impact exposures an understudied research area. Included in this set of low-impact exposures is seasonal variance at birth.

This thesis measures the effects of seasonal variance at birth on lifetime disease risk at both the population-level and molecular-levels. Four aims, comprising this thesis study, were conducted that utilize data from pharmacology, clinical care (Electronic Health Records) and genetics. These aims included: 1.) Development of an Algorithm to Reveal Diseases with a Prenatal/Perinatal Seasonality Component (described in chapter 2); 2.) Investigation of Climate Variables that Affect Lifetime Disease Risk By Altering Environmental Drivers (described in chapters 3 and 4); 3.) Discovery of Genes Involved in Birth Season – Disease Effects (described in chapter 5) and 4.) Investigation of Pharmacological Inhibitors As

Phenocopies of the Birth Season – Disease Effect (described in chapters 6 and 7). Knowledge gained from these four areas, through seven distinct studies, establishes that birth season is a causal risk factor in a number of common diseases including cardiovascular diseases.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank Dr. Nicholas P Tatonetti for being a fantastic mentor, adviser and guide throughout my dissertation studies. Without his enthusiastic support, I would not be here today writing and defending my dissertation. Many times throughout this process I have wanted to give up and throw in the towel, but I have persevered in large part due to your belief in my ability to overcome all of the challenges placed in my path.

I would like to thank all of my committee members: George Hripcsak, Dennis Vitkup, Andrew Gelman and Pierre Gentine. You have taken the time to meet with me to discuss my research, career goals and other related topics and I am forever in your debt. I would also like to thank the training directors and staff from the Precision Medicine training grant who have generously provided training support and guidance to me along the way. Specifically, Henry Ginsberg, Wendy Chung, Siqin Ye, Alex Fedotov, Sarah Oldham, and Sophia Li Ferry. I have enjoyed the lunches, and relaxing chats on all sorts of 'precision medicine' topics. You have livened my time at Columbia University and made it an unforgettable experience.

I would like to thank the Department of Biomedical Informatics who I have had the pleasure to work with these past seven years. I began as a master student and research staff officer back in 2010 and finally as a PhD student from 2013 – 2017. I would like to thank Chunhua Weng for

helping me transition the gap between full-time researcher and ultimately a doctoral student. Suzanne Bakken for her advice and guidance, and for being a shining example of a successful woman in academia. There are so many people to thank in the department, whose roads have crossed mine during my time at DBMI. Including numerous cohorts of masters students and doctoral candidates that I am proud to call my colleagues and friends. The entire Tatonetti Lab has made my life at Columbia University enjoyable, exciting and science-filled. Specifically, I would like to thank Carol Friedman and Marina Bonanno for their work with the National Library of Medicine's informatics training program (which funded my dissertation work for 2 years). I would like to thank Raquel Perez, Kirsy Toribio, Kang Chen, Rosemary Vasquez, Achilles Sanchez, Carol Pitter, Deidre and the entire administrative staff for their tireless work making sure that the department was up and running.

I would like to thank my earliest informatics supporters back at Saint Vincent College (my undergraduate alma mater), who first sparked my interest in the field of Bioinformatics and Computer Science. Specifically, Michael L Sierk, Cynthia Martincic and Mandy Raab. You empowered me to leave the field of biochemistry and move to this exciting interdisciplinary space that is bioinformatics – the ultimate nexus between computer science, biology and medicine.

Finally, I would like to thank my friends and family. My best friends and former college roommates: Briana Taylor Keith, Malori Stevenson and Grettelyn Nypaver Darkley. For the countless hours we have spent chatting, drinking tea and doing non-dissertation stuff, I am forever grateful. For my three grandparents that passed on during the course of my time at Columbia University, I am greatly indebted to each of you more then words can express: Daniel Lennon, Patricia Lennon and Daniel Bernard Boland. *Requies in pace*. I also would like to thank

my surviving grandmother, Brigid Boland, who always encouraged me to push myself beyond what I thought possible. Life would not be complete without my brothers and sister, nieces and nephew / godson, you are an endless source of joy and love. Lastly, I thank my father, Paul J Boland, for providing me with more then life, but also strength when I was weak, encouragement when I was down, joy when I was morose, courage when I was despondent, and food when I was hungry.

Dedication

"Wherever the art of medicine is loved, there is also a love of humanity" ~Hippocrates

We, as humans, struggle through this world bearing our sorrows oftentimes alone. In this light, I have chosen to dedicate my work to all who have lost a child, infant or fetus before his/her/its time. May this work bring humanity one step closer to understanding the unfathomable and explaining the inexplicable.

As a corollary, I would also like to dedicate this work to my namesake patrons: Our Lady Derzhavnaya, Martin de Porres, Raphael the Archangel, and Mary of Egypt. *In lumine Tuo videbimus lumen.* May this research work contribute to the betterment of humanity.

Chapter 1

Introduction

Since antiquity (Hippocrates and Galen, 1952), the relationship between disease and birth seasonality was described, pondered and studied. The Ancient Egyptians, some 5,500 years ago, were known to apply different treatments for a given condition depending on the season of the year, which was noted by specially labeled treatment jars (Allen, 2005). Hippocrates wrote down and recorded tomes of health information from the Ancients. He described a connection between seasonality and disease nearly 2,500 years ago, *"for knowing the changes of the seasons…how each of them takes place, he* [the clinician] *will be able to know beforehand what sort of a year is going to ensue…for with the seasons the digestive organs of men undergo a change*" (Hippocrates and Adams, 460BCE).

Not only was seasonality deemed of vital importance to the Ancients, but also climate factors – wind, water, air quality – were viewed as critical to human health. Hippocrates described the vital connection between climate factors and overall health and surgical outcomes. The position of cities was important because of the city's exposure to these climate factors. He wrote *"cities that are exposed to winds between the summer and the winter risings of the sun and … those*

which lie to the rising of the sun are all likely to be more healthy than such as are turned to the North, or those exposed to the hot winds." (Hippocrates and Galen, 1952)

Childbirth and fertility were a main focus of Ancient medicine with fertility seasons being noted in many Ancient cultures. Traditionally, fertility seasons were often seen as ways to promote the survival of the offspring by planning the child's birth in a mild season (thereby increasing the child's chances for survival). In modern times, the effect of birth season on infant mortality is less pronounced except in some Northern European countries where season of conception is still associated with increased infant mortality (Melnikov et al., 2007). Conception during the summer months in Novosibirsk, Siberia was associated with increased adverse pregnancy outcomes (Melnikov et al., 2007). A June or July conception (the worst observed in their study) would result in a predicted birth month of February or March (i.e., the middle of the winter) (Melnikov et al., 2007). Also some African countries experience season-of-birth increases in infant mortality due to increases in infectious diseases that are driven by the season, including malaria and cholera (Becher et al., 2004). In those countries, being born during the rainy season was associated with increased risk of infant mortality (rate ratio of 1.21) (Becher et al., 2004).

1.1 Literature Review

1.1.1 Epidemiological Review

Prenatal exposure to infectious disease is known to increase the risk of adverse fetal outcomes. The classic example is rubella exposure during pregnancy (Hill, 1965), which results in fetal loss and intellectual / communication problems in offspring born following prenatal rubella exposure (Hardy et al., 1969). A large study of a rubella outbreak in 1964 was conducted on prenatal exposure to rubella and the various outcomes of offspring (Sever et al., 1965). Surprisingly, they found that for first-trimester rubella exposure: 84.4% of conceptions were live-born healthy (519

/ 615), 7.2% (44 / 615) were lost during the prenatal period, 2.0% were born with a congenital rubella syndrome / malformation (12 / 615), and 6.5% were lost to follow-up (40 / 615) (Sever et al., 1965). The rubella outbreak and Sever et al. study was conducted prior to the legalization of abortion in the USA, which occurred in 1973. This explains why elective termination / induced abortion is not present in the statistics for rubella exposure.

According to the Centers of Disease Prevention and Control (CDC), the current background rate of fetal loss during the prenatal period is 17% (CDC et al., 2012; Ventura et al., 2012). This is three times the reported rate of fetal loss experienced by the rubella-exposed mothers in 1964 as reported by Sever et al.. Currently, only 62.6% of conceptions are live-born healthy when all outcomes are accounted for including the 18.4% of conceptions that end in legal termination (CDC et al., 2012; Ventura et al., 2012). This is without any known prenatal exposures. Although with a known exposure (e.g., a pharmacological drug), the risk to the fetus can increase dramatically (Boland and Tatonetti, 2016a).

Recently, the Zika virus has gained notoriety in the media. While not proven, there is some evidence to suggest that prenatal exposure to the Zika virus results in microcephaly in the offspring (Rasmussen et al., 2016). Some researchers believe that the detrimental effects of Zika virus on the offspring may be exacerbated by other confounder variables, e.g., nutrition. No prior outbreaks of Zika virus (in the Pacific Islands) were linked with microcephaly in prenatally exposed fetuses (Duffy et al., 2009; Ioos et al., 2014). This suggests that other factors may be behind the sudden increase in microcephaly in Zika exposed areas. Whatever the true cause of the rise in microcephaly, this recent outbreak does underscore the importance of prenatal exposures on fetal outcomes and the large variability that exists for a given exposure both among the women exposed and also their offspring.

Prenatal exposures resulting in immediate adverse outcomes that are easily observable at birth (e.g., microcephaly) are studied more often in the literature. However, exposures (even high-impact exposures) with effects that occur much later in life are more difficult to study (Boekelheide et al., 2012). Sever *et al.* also noted that while 84.4% of rubella-exposed conceptions were live-born healthy, there was not sufficient time to study the long-term outcomes from prenatal rubella exposure (i.e., a baby may appear healthy at birth but may develop a disease later in life due to rubella exposure) (Sever et al., 1965).

Several low-impact exposures have been linked with lifetime disease risk including arsenic and lead exposure in water. The immediate outcomes following prenatal exposure are easier to capture. The immediate effects following prenatal exposure to arsenic in water included an increased risk of spontaneous abortion (i.e., "miscarriage"), stillbirth and neonatal death (Milton et al., 2005). Others found that prenatal exposure to arsenic resulted in low birth size including reduced head size (i.e., "microcephaly") and chest size (Rahman et al., 2009). However, while lifetime effects from ***prenatal*** exposure to arsenic may be difficult to quantify, the effects from ***general*** exposure on disease risk were easier for researchers to quantify. Arsenic exposure in drinking water has been linked to increased rates of various cancers, including lung, kidney, skin, bladder and liver cancers (Chiou et al., 1995; Hopenhayn-Rich et al., 1998). In Taiwan, researchers found a link between arsenic exposure and cardiovascular diseases (Navas-Acien et al., 2005) and mortality from ischemic heart disease (Chen et al., 1996). However, in other settings and locations the findings were less clear suggesting that the mechanism connecting arsenic exposure and cardiovascular disease may require further elucidation (Navas-Acien et al., 2005).

Another study found that both lead and arsenic exposure in drinking water could affect the

neurodevelopment of children (Calderón et al., 2001). They found that arsenic influenced verbal abilities and long-term memory functioning while lead exposure affects the attention processes (Calderón et al., 2001). Lead exposure was linked with neurological development in children in studies containing data from around the world (Lanphear et al., 2005). Additionally, a causal relationship has been identified between lead exposure and hypertension (Navas-Acien et al., 2007). Hypertension is a common risk factor for severe cardiovascular disease.

Clearly, infectious disease exposure along with exposures to heavy metals during the prenatal period can have a critical impact on fetal outcomes. Many of these factors (especially infectious diseases) vary seasonally, leading modern researchers to investigate the relationship between developmental seasonality (using birth month as a proxy) and disease risk. Several recent studies have linked birth month with neurological (Halldner et al., 2014; McGrath et al., 2010; Willer et al., 2005), reproductive (Huber et al., 2008; Huber and Fieder, 2009; 2011; Huber et al., 2004; Kemkes, 2010), endocrine (Kahn et al., 2009) and immune / inflammatory disorders (Disanto et al., 2012), and overall lifespan (Doblhammer and Vaupel, 2001). Each of these studies uses birth month as a proxy for birth season.

Major radiation exposures – such as Chernobyl and Fukushima-Daiichi – are also studied using various climate models to accurately capture the spread of radioactive particles over large areas (Møller and Mousseau, 2006; Steinhauser et al., 2014). Most models incorporate wind speed and current, and precipitation patterns (Liu et al., 2001). Many of these climate factors are shown to be important when birth month - disease patterns are compared across countries (chapter three of this dissertation) Correlating each birth month – disease pattern with some known climate factors (e.g., precipitation patterns) per region will allow for meaningful comparisons of results. This will reveal the underlying climate factor driving the birth month - disease association.

1.1.2 Genetics Review

1.1.2.1 Learning from High-Impact Exposures: The Importance of Prenatal Regulation of Vitamins

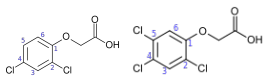
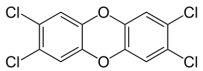
Birth season – disease relationships are thought to exist due to some seasonal change in climate or pollutant exposure, which thereby alters disease risk. The mechanisms underlying these relationships involve some 'gene-environment' relationship (Boland et al., 2013b; Boland and Tatonetti, 2016c; Burga and Lehner, 2012). The field of 'epi-genetics' focuses on elucidating genetic changes that occur due to environmental contaminants (Egger et al., 2004). Typically these changes take the form of DNA methylation changes, or histone modifications, which can be passed on to offspring (Egger et al., 2004). Some traditionally 'genetic' conditions, such as Down's syndrome increased dramatically exactly nine months following the Chernobyl nuclear disaster (Scherb and Voigt, 2007). These increases were observed in Berlin, Germany and Belarus (Scherb and Voigt, 2007) indicating that some chromosomal instability conditions are related to radiation exposure.

A look at several 'high-impact' prenatal exposures reveals a common theme of sex-ratio disturbances. **Table 1** depicts a summary of well-studied exposures and their resulting fetal outcomes. A variety of exposures resulted in sex-ratio disturbances. Data from the Seveso Herbicide plant explosion demonstrated that fathers exposed to dioxin yielded ***more*** female offspring. Contrastingly, fathers exposed to radiation yielded ***less*** female offspring as demonstrated in Hiroshima, Nagasaki, and Chernobyl. The chronic sex disturbances observed in Europe (principally Northern Europe) are believed to be due to abnormalities in sperm production that were observed in wildlife in the surrounding region following the Chernobyl nuclear disaster. Barn swallows were found to have abnormal sperm with observed reductions in

vitamins A and E (Møller et al., 2005). Increased blood oxidation levels of lipids were observed in Chernobyl children with rates higher in girls then in boys (Ben-Amotz et al., 1998). Beta-Carotene (pro-vitamin A) supplementation lowered these oxidation levels in the Chernobyl children leading researchers to conclude that irradiation increases lipids' susceptibility to oxidation and that Beta-Carotene supplementation may act as a lipophilic antioxidant or radioprotector (Ben-Amotz et al., 1998).

Vitamin supplementation can potentially mitigate many adverse effects following exposure to catastrophic events (**Table 1**). Many vitamins have been linked to prenatal outcomes. A lack of folate (vitamin B9) during prenatal development has been causally linked to spina bifida and neural tube defects in the fetus. Increased folic acid (the synthetic form of vitamin B9) supplementation in pregnant women resulted in a reduction in neural tube defects outcomes in offspring (Czeizel and Dudas, 1992). Furthermore, folate is known to vary seasonally in humans as observed in Gambia, and China (Bates et al., 1994; Hao et al., 2003). Therefore, variance in prenatal exposure to various critical vitamins (e.g., folate) resulting from seasonal changes may be responsible for some birth season – disease effects.

Table 1. Summary of High-Impact Prenatal Exposures And Their Effects On Offspring

Disaster Name, Location, Date	Adverse Fetal Outcomes		Refs.
	Immediate	**Prolonged**	
Hiroshima and Nagasaki, Japan, Aug. 6 and 9, 1945	**Neurological damage:** • Severe neurological impairment and / or lowering of intelligence • **Risk**: 7 - 26 weeks after fertilization **Microcephaly:** • **Risk**: 0-18 weeks after fertilization	**Sperm Effect:** • Irradiated fathers yield ***less*** female offspring **Egg Effect (?):** • Irradiated mothers yield less male offspring	(Green, 1968; Otake et al., 1988; Scherb and Voigt, 2007; Schull et al., 1988; Yamazaki and Schull, 1990)
Agent Orange and United States Service Men and Women Serving in Vietnam Area, 1960-1968 	None • No Pregnancies Reported During Vietnam Deployment	**Male or Female** Exposure: • Spina Bifida **Female** Service in Vietnam (not tied to any known herbicide, dioxin or agent orange): • *Cardiovascular* • *Reproductive* • *Neurological* • *Structural*	(VA, 2015a; b)
Seveso Herbicide Plant Explosion, Italy, July 10 1972 	**Spontaneous Abortions:** • 67.7% increase **Chloracne** • Present in children (not a prenatal effect)	**Sperm Effect:** • Fathers exposed to TCDD or dioxin yielded ***more*** female offspring	(Clapp and Ozonoff, 2000; Hay, 1977; Mocarelli et al., 1996; Mocarelli et al., 2000)
Chernobyl, Ukraine, USSR, April 26 1986	**Spontaneous Abortions:** • Increased **Fetal Loss:** • Excess ***loss of 400 male fetuses*** – Czech Republic exposed during end of 1st trimester **Congenital Malformations:** • Increased **Down's Syndrome:** • Peaked in Germany (Berlin) and Belarus, 9 months afterwards	**Sex Ratio Disturbances**: • Throughout Northern Europe **Thyroid cancer:** • Increased **Barn Swallows:** • Experienced abnormal sperm and reductions in Vitamins A and E	(Jacob et al., 1998; Møller and Mousseau, 2006; Møller et al., 2005; Peterka et al., 2004; Scherb and Voigt, 2007)
Fukushima-Daiichi Nuclear Disaster, Japan, March 16, 2011	**Fetal Loss / Anomalies:** • No Observed Change	To Early To Determine	(Fujimori et al., 2014)

1.1.2.2 Understanding the Evolutionary Gradient for Genes that Regulate Seasonally Varying Biofactors

Evolutionary gradients involved in seasonally varying processes have been described in many species including birds (Porlier et al., 2009), insects (Hoffmann and Weeks, 2007), snails (Trussell and Etter, 2001) and various plant species (Etterson et al., 2016; Kay and Sargent, 2009). In some species, these changes have been tied to latitude as well (Schemske et al., 2009). Climate variation has been shown to drive evolution in human genes generally speaking (i.e., without the seasonal aspect) (Piazza et al., 1981). The specific climate-driver involved varies by species. For example, in snails an evolutionary gradient was observed in genes that corresponded to snails adaption to water temperature (Trussell and Etter, 2001).

Two types of evolutionary gradients are described in the literature based on the direction of the change: ***co-gradient variation*** (co-vary in the same direction) and ***counter-gradient variation*** (co-vary in opposing directions) (Trussell and Etter, 2001). The importance of both ***co-gradient variation*** and ***counter-gradient variation*** depends on the particular climate variable of interest. These types of variation can enable insights into populations of patients at increased risk for an adverse prenatal outcome through perturbation of a seasonal mechanism.

1.2 Problem Statement

Developmental defects occur in 100,000 to 200,000 children born each year in the United States of America. 97% of these defects are from unidentified causes. These causes are believed to be due to unidentified prenatal/perinatal environmental exposures. Many outcomes (e.g., developmental defects), result from interactions between genetic and environmental factors. The literature describes a wide-variety of prenatal / perinatal environmental factors and the resulting outcomes on the offspring. However, previous research focuses heavily on high-impact

catastrophic events. These catastrophic events include the Hiroshima and Nagasaki explosions and subsequent radiation exposure, and more recent events such as the Fukushima-Daiichi nuclear disaster. However, the lifetime effects from prenatal exposures with low impact (e.g., lead paint) are often understudied. Even when these lower impact exposures are studied, the focus is often placed on immediate effects of the exposure (e.g., fetal anomalies, miscarriage rates) leaving lifetime effects largely unexplored. Thus, making prolonged (or lifetime) effects of low-impact exposures an understudied research area. Included in this set of low-impact exposures is seasonal variance at birth. A body of literature exists focusing on seasonal exposures at birth and conception and the effects that these seasonal birth exposures can have on risk for diverse diseases including asthma, attention deficit hyperactivity disorder (ADHD) and various reproductive conditions among others. **The studies comprising this dissertation assert that the effect of seasonal variance at birth on lifetime disease risk is an under-studied research area (literature gap) that is worthy of systematic investigation on both population-level and mechanistic-levels**.

1.3 Purpose of the Study

The purpose of this study is to systematically explore the relationship between seasonal exposures at birth (perinatal) and during the gestational period (prenatal) and their disease-related effects later in life. As previously stated, the bulk of prior research focuses on environmental effects from high-impact exposures, such as Hiroshima and Nagasaki. When research on low-impact exposures is conducted, it is typically focused on immediate effects such as miscarriage rates and fetal anomalies. Seasonal variance at birth (a type of low-impact exposure) and the effect on lifetime disease risk (a prolonged outcome) is an understudied research area.

A systems-level approach is needed to understand the seasonal factors driving the relationship between season and lifetime disease risk. Important information can be gleaned from genes involved in evolutionary gradients based on geographic location. These genes have been dictated by evolution as being important in the climate-survival relationship and therefore are likely to be important for key seasonally dependent factors (e.g., Vitamin D).

A conceptual schema detailing the relationship between different factors that affect prenatal / perinatal outcomes following exposures is shown in **Figure 1.** Certain genetic variants exist in evolutionary gradients that are based on the geographic locations of the inhabitants. Information on these genes and the proteins perturbed by these genetic variants along with the immediate prenatal outcomes that result (e.g., miscarriage rates, fetal anomalies) could be utilized to help understand the birth season – lifetime disease risk effects that are observed at the epidemiological level. This is represented in **Figure 1** by the orange in-direct relationship arrow that connects the gene's evolutionary gradient and related associations between season and disease (i.e., prolonged adverse effect).

The goal of this dissertation work is to develop a systems-level approach to understand the seasonal factors of early development with clinical and pharmacological applications. My overall approach is informatics-driven while harnessing methods from Earth Science, Pharmacology, Genetics, and Epidemiology when appropriate.

1.4 Research Questions and Hypotheses

This research consists of three main research questions with their associated hypotheses. The questions are as follows:

Q1: Does a patient's birth month (as a lower-level proxy for season) affect their lifetime disease risk? If so, what types of diseases (i.e., disease categories) correspond with each season?

- H1: Undetected developmental birth defects occurring during the prenatal / perinatal period can manifest themselves later in life as chronic diseases

Q2: What biologically important factors (i.e., biofactors) vary seasonally in humans? Of the known seasonally varying biofactors, which biofactors are expressed during the prenatal period? Which of those biofactors are potential modulators of birth season – disease effects?

- H2: Seasonally varying factors (e.g., vitamin D, folate) during the prenatal period can affect development

Q3: Can birth month be used as a proxy for undetected environmental exposures that affect disease risk? How do these effects vary by climate, region, and country? Are the seasonally varying factors (e.g., low sunlight) correlated with the disease risk experienced at each site/climate/region/country?

- H3: Birth month is a proxy for undetected environmental exposures that affect personal disease risk later in life. These effects are also tied to the climate at birth; therefore being born in October in New York City is very different from being born in Taiwan in October. However, the disease-associated mechanism (e.g., low vitamin D at birth and adjustment disorders) should be generalizable across countries even if the specific month when vitamin D is low varies by country. This holds if and only if the variance in the exposure truly confers disease susceptibility.

To address these questions, this dissertation focuses on four aims. **Table 2** lists each aim, its corresponding study and chapter reference along with relevant publications resulting from the work.

Table 2. Dissertation Aims, Corresponding Studies, and Chapter Reference Information

Aim	Description	Ch.	Ref.
	Section I: Population-Level Insights		
1	*Develop an Algorithm to Reveal Diseases with a Prenatal/Perinatal Seasonality Component*		
	• Season-Wide Association Study (SeaWAS) algorithm development and results from Columbia University Medical Center	2	(Boland et al., 2015b)
	• Validation of novel cardiovascular disease – birth month associations from SeaWAS at Mt Sinai Medical Center	2	(Li et al., 2016)
2	*Investigate How Climate Variables Can Affect Lifetime Disease Risk By Altering Environmental Drivers*		
	• Large-scale exploration of Climate-Hospital Performance Effects Using Medicare's Hospital Compare Dataset (over 4,700 hospitals from 15 climates)	4	(Boland et al., 2017b)
	• Multi-Site / Multi-Country SeaWAS to Find Environmental Drivers Correlated With Birth Month Associations	3	(Boland et al., 2015a; Boland et al., 2017c)
	Section II: Mechanistic Insights		
3	*Uncover Genes Involved in Birth Season – Disease Effects*		
	• Construct Month-of-Birth (MOB) – Gene Map	5	(Boland and Tatonetti, 2016c)
4	*Use Pharmacological Inhibitors As Phenocopies of the Birth Season – Disease Effect*		
	• Systematic exploration of a gene – DHCR7– under evolutionary pressure due to geographic location reveals that pharmacological inhibition results in adverse effects in similar body systems, but much more severe, to those experienced by individuals born in certain months	6	(Boland and Tatonetti, 2016a)
	• Assess whether pharmacological inhibition of identified genes results in adverse outcomes	7	(Boland et al., 2017d)

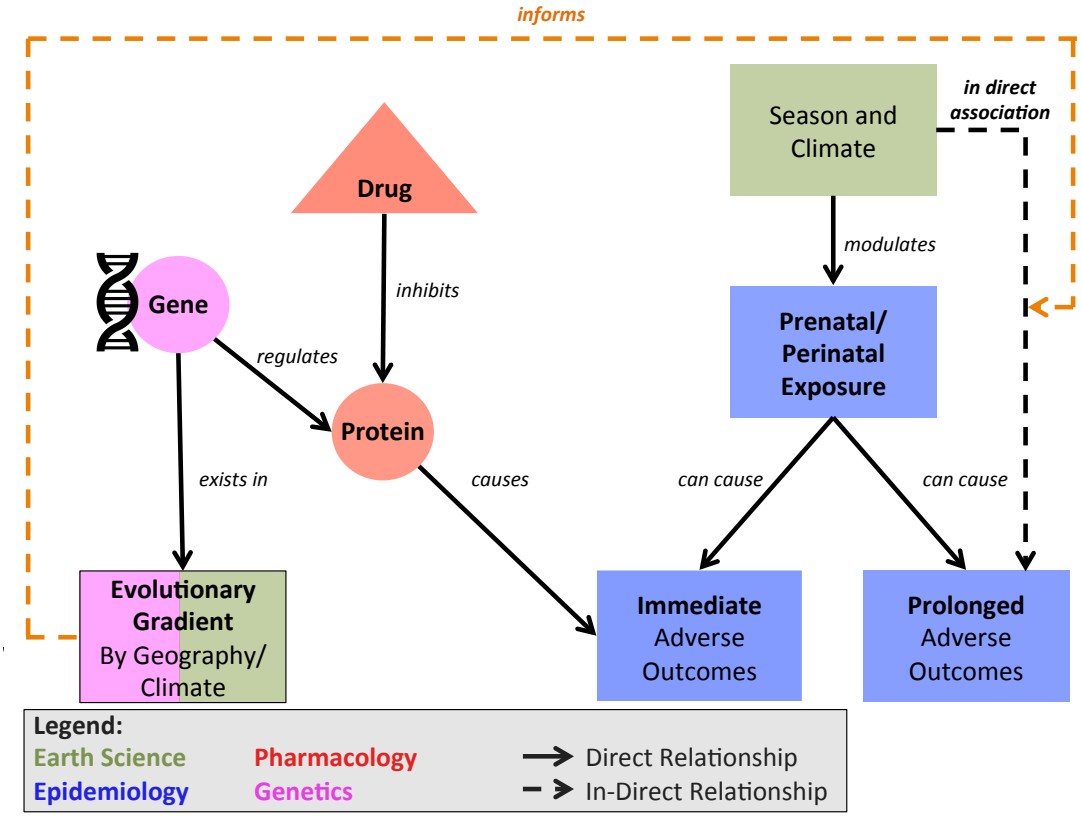


Figure 1. A Conceptual Schema Detailing The Relationship Between Different Factors that Affect Prenatal / Perinatal Exposures Outcomes. This Knowledge Is Integrated In The Studies Comprising This Dissertation To Inform Our Understanding Of Season – Lifetime Disease Risk Relationships.

1.5 Significance

"On relatively small evidence we might decide to restrict the use of a drug for early-morning sickness in pregnant women. If we are wrong in deducing causation from association no great harm will be done. The good lady and the pharmaceutical industry will doubtless survive….but we should need very strong evidence before we made [make] people burn a fuel in their homes that they do not like or stop smoking the cigarettes and eating the fats and sugar that they do like." -Hill 1965

In 1965, Dr. Hill outlined several important steps in determining whether an association was causal or not (Hill, 1965). He also noted an **important facet of causality in epidemiology – to error on the side of increased safety to mothers and their offspring**. This is of vital importance and remains a hallmark to this day. My dissertation seeks to embrace this approach and error on the side of increased safety to mothers and their children while learning as much as possible about the birth season - disease effect to provide insights into the scale of the issues, and the severity of the risk faced by exposed children.

Major developmental defects occur in 100,000 to 200,000 children born each year in the United States of America (USA). However, a much larger concern in the USA remains the children afflicted with learning disabilities such as Attention Deficit Hyperactivity Disorder (ADHD). The number of those afflicted with ADHD continues to grow. Estimates from the Department of Education (DOE) report that 6.4 million children and youth ages 3-21 received special education services representing 13% of all public school students nationally (NCES, 2016). This represents a massive public health issue. In aim one, ADHD was found to be associated with fall birth (November was the peak risk month) in NYC. If this study can identify shed light on the mechanisms behind the ADHD – season of birth issue this could be of great importance to the DOE and funding agencies.

Asthma is another common disease of childhood. In 2014, asthma was responsible for 439,435 inpatient hospital discharges (CDC, 2016). Among children under the age of 18, 8.6% are afflicted with asthma (CDC, 2016). This makes asthma another major public health problem for children. Asthma was found to be linked with season of birth in aim one, which was validated in the literature by a study published in 1983 in Denmark (Korsgaard and Dahl, 1983). The climate driver was peak sunshine exposure, which was related to high temperature and humidity. High

temperature and humidity provide an ideal breeding ground for microscopic dust mites. Exposure to these dust mites during the first few months of birth (i.e., perinatal / post-natal exposure) increases the likelihood that the child will develop dust mite allergies later in life. This also contributes to an increased risk of asthma. Aim two seeks to understand this relationship more deeply by uncovering how disease risk for asthma varies across various sites throughout the world.

Asthma is primarily a disease of the poor - 10.4% of children below the poverty level were afflicted with asthma vs. 6.3% of those in the richest group (i.e., 450% of poverty level or higher) (CDC, 2016). Finding and identifying the key drivers behind the birth season – asthma relationship will be tremendously helpful to this traditionally underserved population.

Additionally, this dissertation seeks to understand not only the drivers of these diseases at the epidemiological level, but also at the genetic level to uncover the mechanistic reasons for these occurrences (e.g., dust mite prevalence, air pollution, sunshine exposure). This will lend critical insights into additional genetic risk factors that may play a role in increasing an individual's risk for an adverse effect of birth season.

1.6 Contributions

This dissertation as a whole asserts that the effect of seasonal variance at birth on lifetime disease risk is an under-studied research area (literature gap) that is worthy of systematic investigation on both epidemiology and genetic levels. The first two aims seek to identify the diseases with birth season/month relationships and the climate drivers (e.g., sunlight, rainfall) responsible for the increased disease risk. These aims focus on providing deep knowledge on the relationship between birth month and disease risk.

The last two aims focus on uncovering the mechanistic drivers of the birth month – disease

relationship. This includes identifying the genes involved in seasonally varying compounds that are important in the birth month disease relationship. Further it seeks to find subsets of the population at increased risk for adverse effects of birth season by investigating genetic variants among hypertension patients. Knowledge can also be gleaned from pharmacological inhibitors of genes and can be used as phenocopies of the birth month - disease relationship. This is especially important for genes known to be involved in an evolutionary gradient tied with climate or geographic region.

This dissertation utilizes a two-pronged approach (epidemiological and mechanistic) to provide a 'deep' understanding into the prolonged outcomes following prenatal/perinatal seasonal exposure. Prior to this research, a literature gap existed for low-impact exposures (such as season) and prolonged outcomes (i.e., lifetime disease susceptibility). Seasonal variance during the prenatal and perinatal period represents an important low-impact exposure that is currently understudied. This dissertation addresses this particular type of low-impact exposure.

Several important steps were outlined by Dr. Hill regarding determining if an association is causal or not (Hill, 1965). These nine criteria became known as the Hill's criteria for causality. The different aims in this study address different criteria outlined by Hill. **Figure 2** illustrates the literature gap that exists in the epidemiology literature (**Figure 2A**) and how each aim addresses a different set of Hill criteria (**Figure 2B**).

This dissertation contributes a thorough in-depth analysis of seasonal factors that affect fetal outcomes upon prenatal or perinatal exposure. Each aim focuses on a different set of Hill's Criteria to identify the factors that are causal in the birth season – disease relationship. By applying this systematic approach, spurious factors will also be identified that cannot be directly tied to a biological mechanism. This allows us to tease out mechanisms that are related to **true**

birth season effects from **spurious** effects. This delineation is useful both to other researchers interested in studying prenatal epi-genetic mechanisms and other outcomes following prenatal exposures.

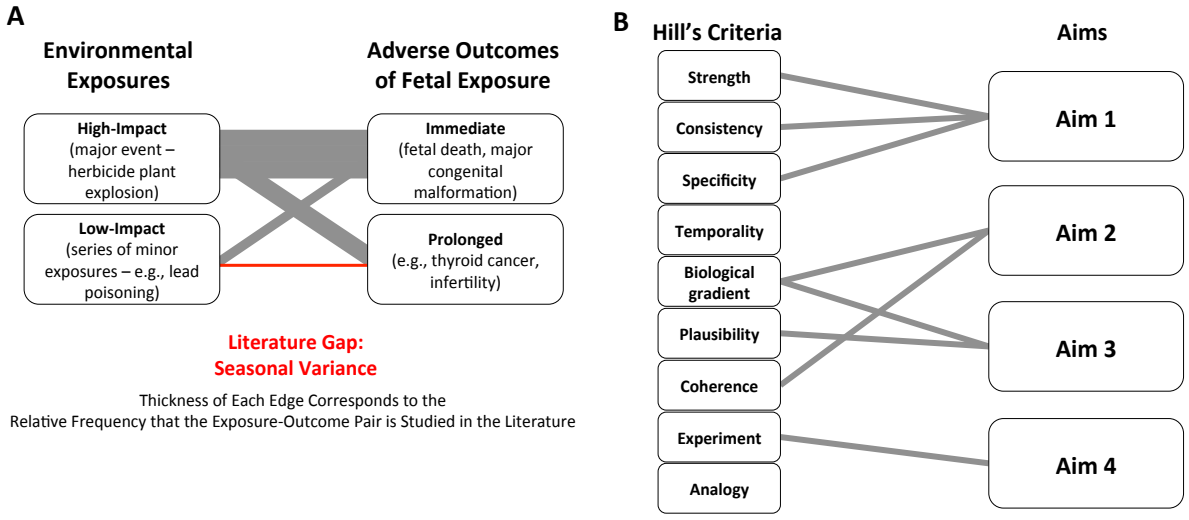


Figure 2. State of the Literature Regarding Environmental Exposures and Their Adverse Outcomes Following Fetal Exposure (Figure 2A) and An Illustration Depicting How Each Aim Addresses a Different Set of Hill Criteria (Figure 2B).

1.7 Overall Limitations

There are several limitations to this dissertation as a whole. Aims one and two (chapters 2-4) rely heavily on data derived from Electronic Health Records (EHRs), which were not collected for research purposes but rather for billing. While a rich data source, there are many biases that exist within EHR data. Data completeness is often an issue, as many patients will have diseases, medications and surgeries that are not recorded in the EHR because those procedures, diagnoses, and so forth were made at another facility. Therefore, patients that are listed as not having a particular diagnosis may in fact have that diagnosis. By testing SeaWAS at multiple sites, many of these biases are minimized. Additionally, Hospital Compare data is used in chapter 4 to assess the relationship between climate and various hospital performance metrics. This contributes valuable information that is free from some of the aforementioned biases.

Aims three and four (chapters 5–7) utilize existing data from PubMed and DisGeNET on disease – gene associations and also seasonally varying biofactors in humans. A lot of this information is incomplete, and represents the various disease selection and publication biases endemic in the scientific community. Many genes are poorly understood and studied with effects that are completely unknown. Additionally, this dissertation focuses on genes that are expressed during development – as annotated by the Gene Ontology. The annotations in the Gene Ontology are also incomplete and this affects the results slightly. Additionally, the fetal-maternal barrier warrants further investigation as the placenta is known to be susceptible to environmental effects (Nelissen et al., 2011). Incorporating knowledge on the epigenetics of the placenta could help with understanding the underlying disease mechanism (Nelissen et al., 2011). For the purposes of this dissertation, I focus on population-level insights gleaned from EHR-data and Hospital Compare data. I couple this with mechanistic insights gleaned from pharmacological data and genetics. However, placenta gene information is not specifically investigated and represents future work for myself or others in the scientific community to explore.

Section I

Population-Level Insights

Chapter 2

Development of an Algorithm for Conducting Birth Season-Wide Association Studies (SeaWAS)

2.1 Abstract

An individual's birth month has a significant impact on the diseases they develop during their lifetime. Previous studies reveal relationships between birth month and several diseases including atherothrombosis, asthma, attention deficit hyperactivity disorder, and myopia, leaving most diseases completely unexplored. This retrospective population study systematically explores the relationship between seasonal affects at birth and lifetime disease risk for 1,688 conditions. I developed a hypothesis-free method that minimizes publication and disease selection biases by systematically investigating disease-birth month patterns across all conditions. The initial dataset includes 1,749,400 individuals with records at New York-Presbyterian/Columbia University Medical Center born between 1900-2000 inclusive while the external replication site - Mount Sinai Hospital - included 1,169,599 patients. I modeled associations between birth month and 1,688 diseases using logistic regression. Significance was

tested using a chi-squared test with multiplicity correction. I found 55 diseases that were significantly dependent on birth month. Of these 19 were previously reported in the literature ($p<0.001$), 20 were for conditions with close relationships to those reported, and 16 were previously unreported. Distinct incidence patterns across disease categories were observed. Individuals born in birth months with higher cardiovascular disease incidence (February-June) were also associated with decreased life expectancy in the literature corroborating the findings. Neurological diseases, pregnancy conditions and asthma associations revealed by my method were validated by European studies in the literature. Novel cardiovascular conditions associated with birth month were validated externally using Mount Sinai Hospital data. Overall, individuals born in May and July had the lowest overall disease risk. Lifetime disease risk is affected by birth month and seasonally dependent early developmental mechanisms may play a role in increasing lifetime risk of disease.

2.2 Introduction

The recent adoption of Electronic Health Records (EHRs) allows meaningful use (Jha, 2010) of data recorded during the clinical encounter for high-throughput or 'phenome-wide' exploratory analyses (Boland and Tatonetti, 2015; Denny et al., 2010; Jensen et al., 2012; Roque et al., 2011; Shah, 2013). Using EHR data requires overcoming problems with definition discrepancies (Boland et al., 2013b), data sparseness, data quality (Weiskopf and Weng, 2013), bias (Hripcsak et al., 2011), healthcare process effects (Hripcsak and Albers, 2013) and privacy issues (Loukides et al., 2010). Informatics methods can overcome these challenges, for instance: standardized ontologies minimize definition discrepancies (Elkin et al., 2006), concordance measured across integrated datasets allows for data sparseness and quality assessment (Weiskopf and Weng, 2013), and statistical methods can minimize bias and healthcare process effects

(Dickersin, 1990; Hripcsak et al., 2007; Stern and Simes, 1997). Using informatics approaches, EHR discovery methods (Jensen et al., 2012) were developed with successful applications in diverse areas including: dentistry (Boland et al., 2013a), genetics (Crawford et al., 2014; Denny et al., 2010; Kohane, 2011), and pharmacovigilance (Haerian et al., 2012; Wang et al., 2009). Novel disease association patterns (Doshi-Velez et al., 2014; Holmes et al., 2011), drug-drug interactions (Tatonetti et al., 2012) and seasonal dependencies (Cohen et al., 2014; Melamed et al., 2014; Randolph, 2014) have also been established using EHRs.

Previous birth month – disease studies investigated the relationship between one single disease and birth month. They did not employ high-throughput or 'phenome-wide' approaches (Denny et al., 2010). In many studies, birth month was used as a lower-level proxy for season. However, some studies used the season variables themselves converting the month-day attributes to a birth season variable (Pantazatos, 2014), or in some cases binning months (e.g., Jan-Mar = winter). The most common statistical method used by prior studies was chi-square analysis (35.8%, 19/53) followed closely by regression. However, the countries conducting these studies were more diverse. Only 26.4% of studies were conducted in the United States of America (i.e., 14/53). These statistics are presented in **Table 3.** Countries with only one study were not included in **Table 3.** When grouped together, Northern European countries in general (e.g., Germany, Denmark, Ukraine, Sweden) were responsible for 15 studies or 28.3% of the sample – greater than any other region in the world.

Methods that were seldomly used include Cosinor analysis, and various forms of temporal analysis including time series spectral analysis and Knox clustering (Kulldorff and Hjalmars, 1999). The prior methods focused on already established birth month – disease relationships and the developed models were tailored to those specific diseases using additional model parameters.

The researchers sought to tease out some of the nuances within their datasets to establish what drivers were important in the birth month – disease mechanism.

Table 3. Statistics for Birth Month Studies Methods and Locations: Prior to SeaWAS

Method	Number of Studies Using Statistical Method (N=53 studies)	Location	Number of Studies At Given Location (N=53 studies)
Chi-square	19	USA	14
Regression	14	Sweden	5
ANOVA	6	Australia	3
Ratios	4	Japan	3
Binomial p	1	Israel	2
Cosinor analysis	1	Spanish Language	2
Temporal Analysis:			
Knox clustering	1	Turkey	2
Time series using spectral analysis	1	United Kingdom	2

Unlike prior methods that focused on one disease at a time, SeaWAS is a high-throughput method that explores all diseases' potential association with birth month at a given location. SeaWAS investigates all diseases without preselecting certain "interesting" diseases to be robust to disease selection bias (i.e., only studying 'popular' diseases) and publication biases (Dickersin, 1990; Easterbrook et al., 1991; Stern and Simes, 1997; Vawdrey and Hripcsak, 2013).

This dissertation focuses on data-driven approaches that seek to learn birth month – disease relationships from the data in a 'phenome-wide' manner across a wide-variety of diverse diseases (Denny et al., 2010). In this study, I developed a high-throughput, hypothesis-free algorithm that mines for disease-birth month associations across millions of records. The approach is called **SeaWAS**: **Sea**son-**W**ide **A**ssociation **S**tudy as it finds all conditions associated

with birth month. I show that SeaWAS detects diseases with seasonal components related to early development. The novel cardiovascular disease findings were then validated at a separate EHR in NYC (same city, demographics and climate).

2.3 Methods

2.3.1 Population

I used Columbia University Medical Center (CUMC)'s health record data, previously converted to the standardized Common Data Model (CDM) developed by the Observational Medical Outcomes Partnership (now the Observational Health Data Sciences and Informatics, OHDSI) (Overhage et al., 2012). CUMC data was initially recorded using *International Classification of Diseases, version 9* (ICD-9) codes. These ICD-9 codes were mapped to *Systemized Nomenclature for Medicine-Clinical Terms* (SNOMED-CT) codes according to the CDM v.4 (Overhage et al., 2012). SNOMED-CT captures more clinical content then ICD-9 codes (Campbell and Payne, 1994) making it ideal for phenotype classification. Additionally, using a standardized CDM increases the portability of the method across institutions enhancing data sharing (Margolis et al., 2014).

All individuals were extracted if they were born between 1900-2000 inclusive (N=1,749,400 individuals) and were treated at CUMC (between 1985-2013), demographics given in **Table 4**. The median age of the population was 38 years (interquartile range IQR: 22, 58). I performed a fisher-exact test between the birth month distributions for each sex vs. the average birth month distribution. Likewise the birth month distributions by birth decade (e.g., 1900-1909, 1990-1999) were compared to the overall average birth month distribution. No statistically significant differences were found ($p=1$ for all comparisons). Therefore, yearly and sex-based variation in the birth month distribution is minimal and should not affect the analyses.

Monthly birth rate data for CUMC was consistent with known New York City (NYC) births using data from the Centers of Disease and Control (CDC) for 1990-2000 inclusive (CDC, 2014b). CUMC data were highly correlated with CDC birth rates from Bronx (r=0.833, p=0.001), New York (r=0.796, p= 0.002) and Queens (r=0.791, p=0.002) counties. I performed this verification check because confirming the place of birth for individuals can be complex (Duncan et al., 2014), and was not possible for the CUMC dataset. Subsequently, for the 1990-2000 period I was able to obtain data regarding the number of babies admitted to CUMC on the day of their birth for the 1990-2000 period and found that the proportion (no. of patients admitted to CUMC on their day of birth / no. of patients included in SeaWAS) ranged from 17.97%-31.28% by birth year with the average proportion being 22.98%. CUMC's Institutional Review Board approved this study.

2.3.2 Algorithm

My method investigated associations with birth month across all recorded conditions. A condition was defined as any SNOMED-CT code mapped using the CDM (Overhage et al., 2012). For controls, individuals were randomly sampled from the same EHR population without the disease ensuring that the control sample size was 10 times the size of the case population. The association between birth month (as an integer) and each condition was modeled using logistic regression with significance assessed using chi-square (R v.3.1.0). Therefore, the monthly birth rate was compared between the case and control populations for each condition adjusting for monthly birth month variation effects. For multiplicity correction, only conditions passing the Benjamini-Hochberg adjustment that controls for the false discovery rate (FDR) were selected (Benjamini and Hochberg, 1995). To ensure sufficient sample size across all 12 months,

only conditions having at least 1000 individuals born between 1900-2000 inclusive were included in the study (i.e., 1,688 distinct conditions).

To evaluate SeaWAS, all articles from PubMed with the term "birth month" were extracted along with an additional article referenced by a located article (N=156). I manually reviewed all abstracts and removed articles related to non-humans (N=8), breeding (N=7), sports (N=10), or where birth month was used for another purpose, e.g., for matching controls (N=34), perspective/meta-analysis papers (N=2), papers not available in English (N=2), and one paper with a statistical error noted in PubMed. This process identified 92 relevant articles. I then manually classified each paper by the disease studied and whether they found or failed to find an association. Some conditions associated with birth month in the literature, e.g., height, were not extractable from the EHR (36 diseases were not extractable). In total, 19 diseases reported in the literature could be mapped to EHR conditions. Of those diseases, 16 were positively associated (>50% of literature supported an association) and 3 were not associated (<=50% of literature failed to find an association). I extracted all relevant EHR codes for each of the 16 positive associations (N=172 codes). These literature associations were used for quality assessment of SeaWAS results.

I used the following internal evaluation technique to evaluate novel associations discovered by SeaWAS. I ran the SeaWAS algorithm on a restricted sample comprising 80% of the original sample, randomly chosen. Results were corrected for multiplicity using the Benjamini-Hochberg adjustment that controls the FDR. I took all novel associations (i.e., not reported in the literature) revealed in the restricted sample, and then validated them using the validation set (containing 20% of the original population). 12 of the 16 discovered associations were validated in this manner.

Permutation analysis was also used for empirical evaluation of SeaWAS. My algorithm randomly selected 55 diagnosis codes from the set of 1,688 codes included in the study. The algorithm then set all codes in this randomly derived set as 'positive' associations. Next, the number of positive literature results in each random sample was measured. This was done for 1,000 random samples. The overall distribution of these random samples was compared to SeaWAS results. This allowed for the assessment of the true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV) and the total number of confirmed literature associations obtained from SeaWAS.

For all significant associations, my algorithm calculated the proportion of individuals having the condition using their birth month and day out of all individuals with the same birth month and day. This generated a set of proportions for every day in the year (366 days). My algorithm then used a 2-month window (Kahn et al., 2009) to smooth the daily proportion rate (1 month before the date and one month after the date). The weekly and monthly averages were then computed. An overview of the algorithm is shown in **Figure 3.**

All SeaWAS results were compared to the literature in a binary manner to ascertain if the association was previously reported. Afterwards, the disease-birth month risk plots from the literature were analyzed. I used three criteria to select studies, namely: 1) published raw data; 2) raw data includes some adjustment for natural variation in birth month depending on study region; and 3) disease-birth month data were at a similar granularity level to allow for effective comparisons (e.g., this criterion would exclude studies that grouped multiple diseases together or removed certain disease subtypes). I sought to include pattern data for at least one study per disease category to compare with SeaWAS.

Table 4. Demographics of Patients Included in SeaWAS: CUMC and Mt Sinai

Demographic	CUMC N (%), N=1,749,400	Mt Sinai M (%), M=1,169,599
Sex [1]		
Female	956,465 (54.67%)	678,717 (58.03%)
Male	791,534 (45.25%)	490,600 (41.95%)
Other/Unidentified	1,401 (0.08%)	282 (0.02%)
Race		
White	665,366 (38.03%)	424,803 (36.32%)
Other [1]	456,185 (26.08%)	165,423 (14.14%)
Unidentified / Unknown	386, 533 (22.10%)	256,819 (21.96%)
Black	189,123 (10.81%)	166,950 (14.27%)
Declined	29,747 (1.70%)	NA
Asian	20,746 (1.19%)	45,596 (3.90%)
Native American / Indian	1,511 (0.09%)	2,447 (0.21%)
Pacific Islander	189 (0.01%)	1,094 (0.09%)
Hispanic / Latino	NA	106,467 (9.10%)
Ethnicity		
Non-Hispanic	590,386 (33.75%)	761,535 (65.11%)
Unidentified	458,071 (26.18%)	208,899 (17.86%)
Hispanic	361,123 (20.64%)	199,165 (17.03%)
Declined	339,820 (19.42%)	NA
Other Attributes		**Median (1st, 3rd Quartile)**
Total SNOMED-CT codes per patient	6 (1, 32)	7 (3, 22)
Distinct SNOMED-CT codes per patient	3 (1,8)	5 (2, 10)
Age (year of service – year of birth)	38 (22, 58)	53* (36, 66)
Treatment Year Range	1985-2013	1979 - 2015

[1] Other (includes Hispanics not otherwise identified)

* Computed in days, age in years = age in days / 365.25

2.3.3 Mount Sinai Replication Methods

To perform a proper replication study, the original SeaWAS study was followed as closely as

possible (Boland et al., 2015b). EHR data was obtained from Mount Sinai Hospital (MSH) located in NYC. Both CUMC and MSH are urban medical centers located in NYC and were expected to be subject to similar climate conditions (Boland and Tatonetti, 2016c). EHR data at MSH is represented using a different schema and data model than was used in the original SeaWAS study at CUMC. Therefore the locally obtained International Classification of Diseases, version 9 (ICD-9) codes used at MSH were mapped to the Systemized Nomenclature for Medicine-Clinical Terms (SNOMED-CT) using the mapping table from the CDM v.4 (Overhage et al., 2012). Approval for this study was obtained from the Institutional Review Board at MSH.

All individuals born between 1926 and 2000 inclusive (1,169,599 patients) who were treated at MSH (between 1979 - 2015) were included in this replication study, demographics given in **Table 4**. The median age of the MSH population was 53 years (interquartile range, IQR: 36-66), which skews older than the original CUMC population (median=38 years, IQR: 22-58). Race and ethnicity demographics are represented slightly differently between the MSH and CUMC datasets. Overall sex, race, and ethnicity distributions did not differ significantly between the two institutions ($p > 0.05$, **Table 4**).

The CUMC-only SeaWAS study found novel cardiovascular condition-birth month associations (Boland et al., 2015b), therefore this was the focus of the MSH replication. Also at both institutions (CUMC and MSH) essential hypertension (associated with birth month at CUMC) was the most prevalent disease, signifying the clinical importance of this association. Therefore, only circulatory system conditions (as defined using the ICD-9 codes) that were present in both the MSH and CUMC datasets (i.e., having at least 1000 patients at both MSH and CUMC) were selected. This represented a set of 108 conditions.

In the MSH replication, the SeaWAS algorithm was modified slightly from the public domain (code available here: https://github.com/maryreginaboland/SeaWAS) to fit the database schema of MSH. First a phenome-wide exploration of all birth month – disease associations for conditions with at least 1000 patients (1433 conditions) was performed. I employed the Benjamini-Hochberg method to correct for multiple hypotheses and control for the false discovery rate (FDR) (Benjamini and Hochberg, 1995) similar to the original study.

Next, I performed a post-hoc analysis, using Pearson's correlation, to compare the birth month - disease risk curves between the two institutions (CUMC, MSH). I assessed statistical significance by comparing the actual correlation between the two institutions to an empirically derived null distribution. My algorithm generated empirical null distributions for each condition by randomizing the birth month–disease risk curve from MSH. It then computed the Pearson correlation for the random curve and CUMC's birth month-disease risk curve. This procedure was repeated 1000 times producing 1000 random correlation results. The empirically derived p-value was determined as the proportion of random correlations greater than the actual correlation between true MSH data and true CUMC data divided by 1000. If this value is zero then the p-value is reported as "$p < 0.001$."

To place the findings in the context of biological mechanisms deemed important in birth month associations, I obtained peak flu season data obtained from the Centers for Disease Prevention and Control (CDC) data on flu activity from 1982-83 through 2013-14 (CDC, 2014a). I also compared the serum vitamin D levels reported in Meier et al. 2004 (Meier et al., 2004).

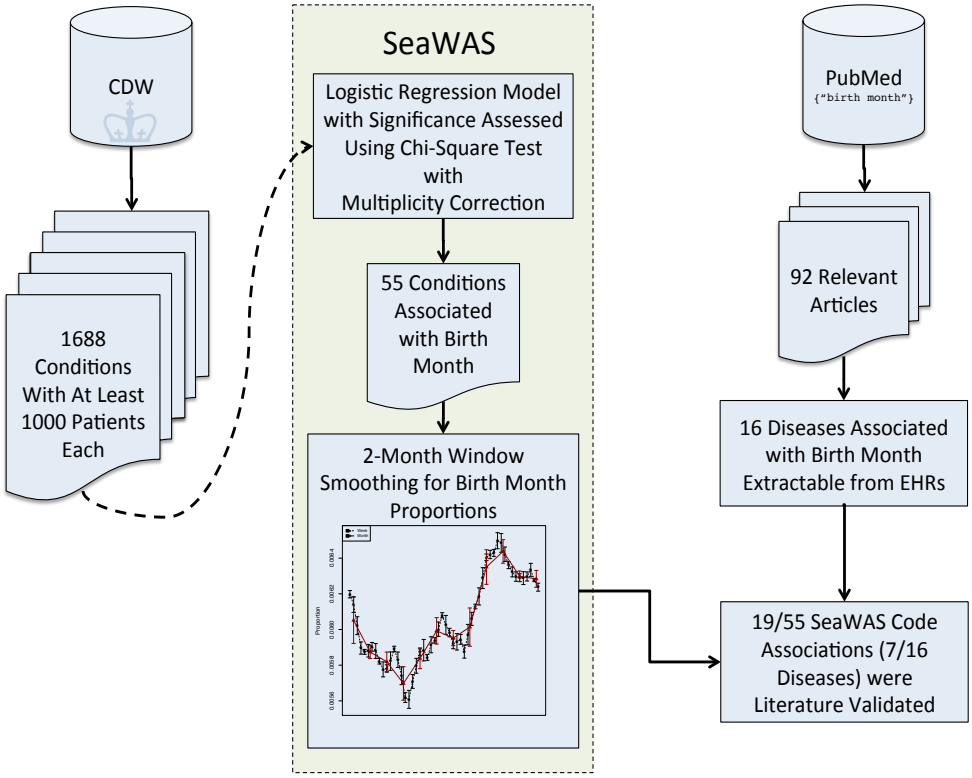


Figure 3. Schematic Overview of Season-Wide Association Study Method. SeaWAS was performed using the Common Data Warehouse (CDW) at Columbia University Medical Center (CUMC). 1,688 conditions had at least 1000 patients and were tested for possible birth month associations. Logistic regression was performed to assess the relationship between birth month (modeled as a numeric variable) and disease. 55 conditions were associated with birth month after false discovery rate correction. Results were compared to relevant PubMed retrieved articles on birth month.

2.3.4 Analysis of Potential Confounders in Cardiovascular – Birth Month Findings

My algorithm revealed a novel relationship between cardiovascular diseases and birth month. Since this finding was novel, I decided to investigate the relationship between the cardiovascular disease – birth month relationship and other known demographic confounders, including ethnicity and income (Cooper, 2001). Therefore, I analyzed the 'Essential Hypertension' (the most common cardiovascular disease - birth month association) in greater depth. Specifically I investigated the relationship between ethnicity and income and the birth month – disease relationship. Patients were separated into 'Poor' and 'Rich' using the IRS 2012 zip code income information (IRS, 2015). Patients were classified as being 'Poor' if the most popular income bracket for their zip code was the lowest IRS income bracket ($1 - $25,000) and patients were classified as being 'Rich' if the most popular income bracket was the highest reported ($200,000 or more) for their zip code.

2.4 Results

2.4.1 EHR mining of 1,688 conditions reveals 55 conditions dependent on birth month

I used SeaWAS to mine birth month associations for 1,688 SNOMED-CT conditions with at least 1,000 individuals recorded at CUMC. After multiplicity correction using FDR (α=0.05, N=1688 conditions), 55 conditions were found associated with birth month. All reported p-values are FDR adjusted (q-values).

2.4.2 Literature validation of SeaWAS results

Using the curated reference set of 16 conditions (that mapped to 172 SNOMED-CT codes), I found 19 SeaWAS results (7 distinct diseases) were supported by the literature, representing a significant enrichment with OR=3.4 (95% CI: 1.9-6.0, $p<0.0001$**).** SeaWAS successfully ruled-

out associations between birth month and disease risk for all 'true negatives' in the reference set**.** The algorithm compared SeaWAS results for known and closely related diseases to help elucidate gaps in the literature. Some diseases, e.g., reproductive performance, featured prominently in both the literature and SeaWAS results, whereas, other diseases featured heavily in the literature but not as strongly in my results, e.g., asthma/allergy and rhinitis. A potential literature gap exists for respiratory syncytial virus (2 publications, **Figure 4),** which had many SeaWAS known or highly related associations (8 total associations). A Manhattan plot visualizes results by disease category showing that some categories including, circulatory, and respiratory diseases appear prominently in the results**.**

The algorithm found 20 conditions associated with birth month that were similar to those in the reference set and 16 that were completely novel (**Table 5**). Nine of these 16 associations were cardiovascular conditions including: atrial fibrillation ($p<0.001$), essential hypertension ($p<0.001$), congestive cardiac failure ($p<0.001$), angina ($p=0.001$), cardiac complications of care ($p=0.027$), mitral valve disorder ($p=0.024$), pre-infarction syndrome ($p=0.036$), cardiomyopathy ($p=0.009$), and chronic myocardial ischemia ($p=0.022$). Seven discovered associations were non-cardiovascular: primary malignant neoplasm of prostate, malignant neoplasm of overlapping lesion of bronchus and lung, acute upper respiratory infection, non-venomous insect bite, venereal disease screening, bruising and vomiting.

2.4.3 Internal evaluation of discovered associations

All novel associations found using SeaWAS were internally validated. The SeaWAS algorithm was run on a dataset containing an 80% restricted sample and then validated using the novel associations in the validation set (20% original sample size). 12 of the 16 novel associations were validated including 6 out of 9 novel cardiovascular conditions. **Table 5** denotes the discovered

conditions that passed the internal validation. Four conditions were not significant after correction in the restricted sample including: mitral valve disorder, pre-infarction syndrome, chronic myocardial ischemia, and vomiting.

2.4.4 Evaluation using permutation analysis

The algorithm uses permutation analysis to assess the concordance between SeaWAS results and what was reported in the literature. It randomly selects 55 codes from the set of 1,688 codes included in the study and set them as 'positives'. The algorithm then measured the number of positive literature results in random samples and compared to SeaWAS. This was run for 1,000 random samples. SeaWAS consistently and significantly ($p < 0.001$) outperformed random for TPR, FPR, and PPV at finding more literature validated associations.

2.4.5 SeaWAS replicates established birth month trends: Asthma, Reproductive Performance and ADHD

I calculated smoothed birth month proportions for all 55 SeaWAS birth month associations. I then compared conditions with known associations to birth month and their published trends. The smoothed weekly and monthly proportions are shown in **Figure 5** for three established associations: asthma, ADHD and reproductive performance and three discovered associations: atrial fibrillation, mitral valve disorder and chronic myocardial ischemia. To compare results with the published proportions from other studies, I used an asthma study from Denmark (Korsgaard and Dahl, 1983), reproductive performance study from Austria (Huber et al., 2004) and an ADHD study from Sweden (Halldner et al., 2014).

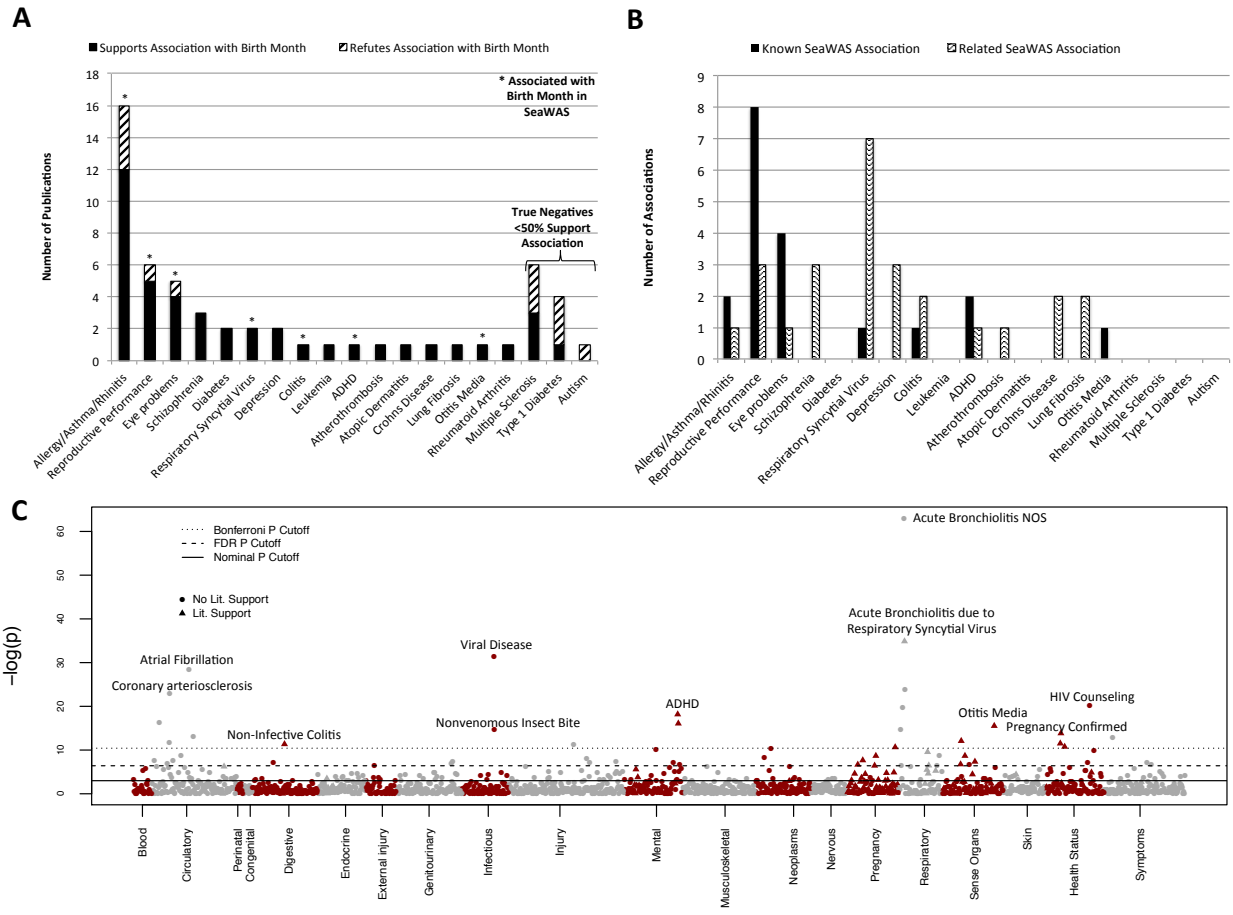
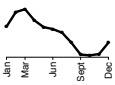
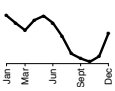
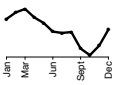
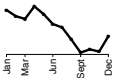
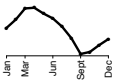
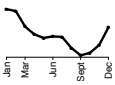
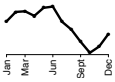
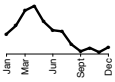
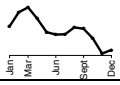
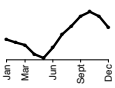
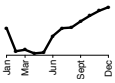
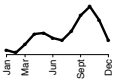
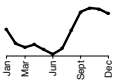
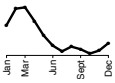
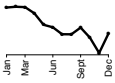
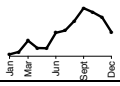


Figure 4. SeaWAS Results Show Enrichments for Literature Associations. Figure 4a shows the breakdown of SeaWAS results by number of publications demonstrating a relationship. **Figure 4b** shows the number of SeaWAS associations known to be related to disease from the literature (solid black), and those that are closely related to known diseases (curvy lines). **Figure 4c** depicts all birth month-disease associations in a Manhattan plot organized by their respective ICD-9 disease categories (x axis). A significant SeaWAS association is a disease-birth month association remaining significant after FDR adjustment.

Table 5. Birth Month-Disease Associations Discovered Using SeaWAS (N=16)

EHR Condition in SeaWAS	N	Passed Internal Validation?	Adjusted P[1]	Seasonal Pattern	Birth Month Risk	
					High	Low
Cardiovascular (N=9)						
Atrial fibrillation	48961	Yes	<0.001		Mar	Oct
Essential hypertension	269913	Yes	<0.001		Jan	Oct
Congestive cardiac failure	61448	Yes	<0.001		Mar	Oct
Angina	20741	Yes	<0.001		Apr	Sept
Cardiac complications of care	13653	Yes	0.027		Apr	Sept
Cardiomyopathy	17873	Yes	0.009		Jan	Sept
Pre-infarction syndrome	25028	No	0.036		Jun	Oct
Chronic myocardial ischemia	10010	No	0.022		Apr	Nov
Mitral valve disorder	22966	No	0.024		Mar	Nov
Other (N=7)						
Acute upper respiratory infection	112487	Yes	<0.001		Oct	May
Bruising	8904	Yes	0.015		Dec	Apr
Nonvenomous insect bite	7435	Yes	0.001		Oct	Feb
Venereal disease screening	69764	Yes	0.003		Oct	Jun
Primary malignant neoplasm of prostate	20353	Yes	0.002		Mar	Oct
Malignant neoplasm of overlapping lesion of bronchus and lung	2714	Yes	0.014		Feb	Nov
Vomiting	30495	No	0.029		Sept	Jan

[1] P-values adjusted using Benjamini-Hochberg method (see Methods)

Comparing results with Denmark's asthma study (Korsgaard and Dahl, 1983) showed highly similar seasonal patterns. They found two large peaks in May and August, with two smaller peaks in June and July (Korsgaard and Dahl, 1983). My results were shifted by 2-months with large peaks in July and October and smaller peaks in August and September. I extracted data on the average monthly sunshine exposure for NYC and Denmark (2014a; 2014b) for comparison (**Figure 5**). For reproductive performance, I compared the results to an Austrian study (Huber et al., 2004) (**Figure 5**). I validated a dip in births among females born in May through September as this was also found in the Austrian study. I compared the ADHD smoothed proportions to odds ratios reported by a Swedish study and found a similar upward trend towards the later part of the year peaking in November (Halldner et al., 2014) (**Figure 5**).

I sought to include at least one seasonality comparison for each disease category (N=7) of known associations to those found by SeaWAS. This includes: allergy/asthma/rhinitis, reproductive performance, ADHD, eye conditions/problems, respiratory syncytial virus, otitis media and colitis. Literature studies on eye conditions/problems failed the three criteria for inclusion as data was presented at different disease granularity levels (e.g., mild myopia was excluded) preventing effective comparisons. I found data for conditions in the three remaining categories, otitis media, colitis and respiratory syncytial virus (**data not shown**). I found many similarities among these data, but the exact mechanistic relationship between these conditions and birth seasonality remains obscure.

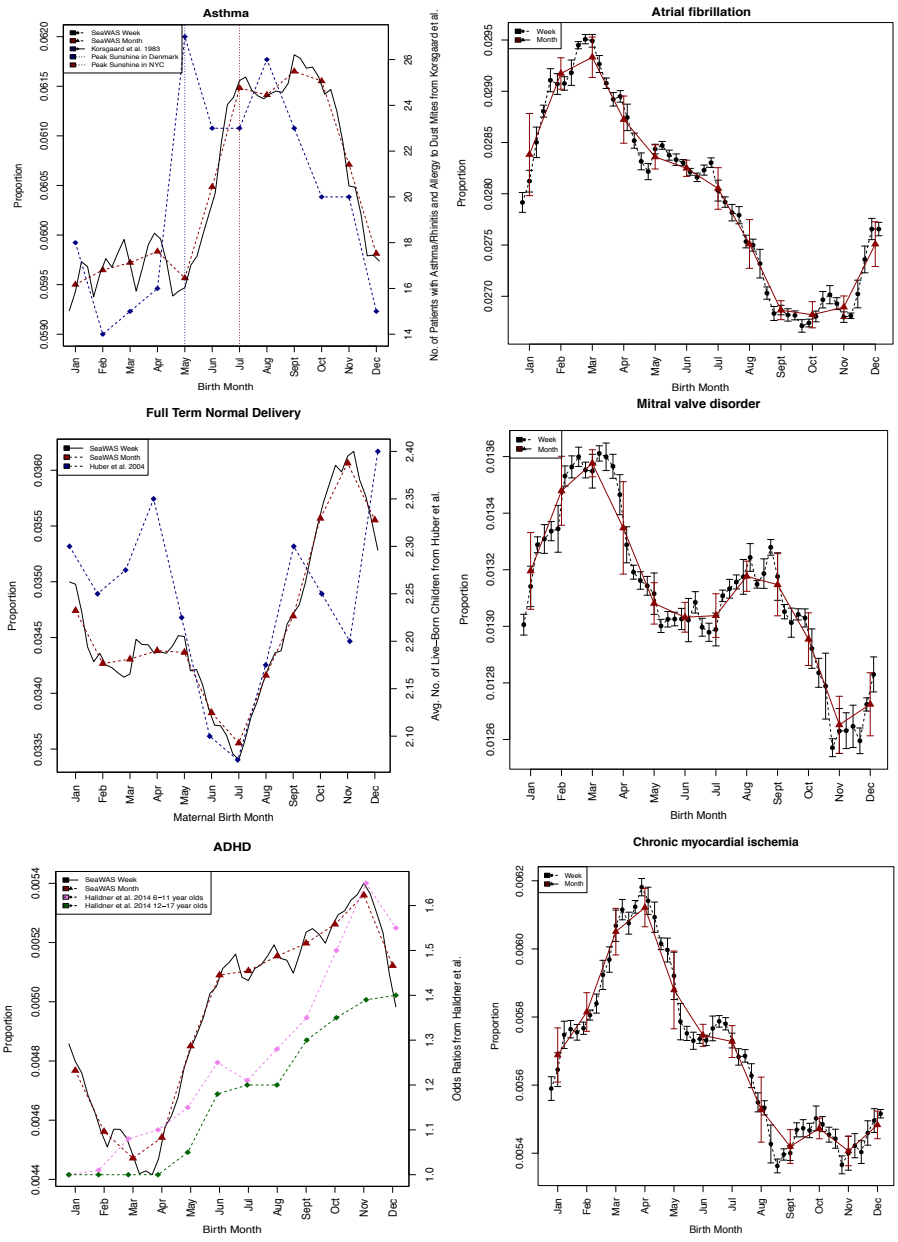


Figure 5. Birth Month Distribution Plots for Three Literature Validated SeaWAS Results and Three Discovered SeaWAS Associations. I selected 3 well-known literature associations: asthma, ADHD and reproductive performance to compare with SeaWAS birth month trends. I compared my results to findings published in articles for each of these diseases: 1) asthma data from a Denmark study by Korsgaard et al. 1983 (Korsgaard and Dahl, 1983); 2) reproductive performance data from an Austrian study by Huber et al. 2004 (Huber et al., 2004), which I compared to Full-Term Normal Delivery (i.e., general birth code); and 3) ADHD data from a Swedish study by Halldner et al. 2014 (Halldner et al., 2014). To facilitate comparison between asthma studies from different locales, I used data on the average monthly sunshine exposure for New York, USA and Skagen, Denmark obtained from World Weather and Climate Information (2014a; 2014b). I also found three interesting new associations: atrial fibrillation, mitral valve disorder and chronic myocardial ischemia.

2.4.6 Discovered Associations: Cardiovascular Conditions and Birth Month

My algorithm found 16 birth month – disease associations with no prior literature, I highlight three of these in **Figure 5**, including: atrial fibrillation, mitral valve disorder and chronic myocardial ischemia. For illustration purposes, I selected cardiovascular conditions whose pattern of association between birth month and disease risk differs. Mitral valve disorder demonstrates a clear bimodal seasonal pattern with a major disease risk peak among those born in March and a second smaller disease risk peak for those born in August. Whereas, risk for atrial fibrillation is unimodal and peaks among those born in March with a trough between September and November.

2.4.7 Patterns of birth-month dependencies cluster by disease type

Of nine discovered cardiovascular associations, six had high-risk birth months in March or April suggesting that high-risk birth months may cluster by disease category. I examined the disease category-birth month relationship and found that individuals born in March were at increased risk for cardiovascular diseases, but they had greater protection against respiratory illnesses and neurological conditions. Contrastingly, individuals born in October were at increased risk for respiratory conditions with increased protection against developing cardiovascular conditions. Overall, some months, namely May and July, had zero at risk diseases.

2.4.8 Cardiovascular Disease Risk-Birth Month And Lifespan-Birth Month

I compared the cardiovascular disease findings (N=10) from SeaWAS to published data relating overall lifespan and birth month (Doblhammer and Vaupel, 2001), see **Figure 6**. Months with lower cardiovascular disease risk corresponded with months having longer life expectancies from Doblhammer *et al.*'s previous study (Doblhammer and Vaupel, 2001). Six of the 10

cardiovascular conditions were significantly anti-correlated with life-expectancy data. The strongest anti-correlation was cardiac complications of care (Denmark: r=-0.815, p=0.001; Austria: r=-0.863, p<0.001); followed by chronic myocardial ischemia (Denmark: r=-0.810, p=0.001; Austria: r=-0.826, p<0.001); pre-infarction syndrome (Denmark: r=-0.712, p=0.009; Austria: r=-0.918, p<0.001); coronary arteriosclerosis (Denmark: r=-0.617, p=0.030; Austria: r=-0.773, p=0.003); atrial fibrillation (Denmark: r=-0.615, p=0.033; Austria: r=-0.763, p=0.004); and angina (Denmark: r=-0.611, p=0.035; Austria: r=-0.771, p=0.003).

2.4.9 Cardiovascular Result Replication at Mount Sinai Hospital

In addition to validating known birth month disease relationships, SeaWAS revealed nine circulatory system conditions were associated with birth month. Specifically winter births (Jan-Mar) were related to increased risk of developing circulatory system conditions later in life. Because this finding was novel, I sought to validate the findings externally using another EHR system also in New York City (NYC). In this way, I was able to test if the relationship between circulatory system conditions and birth month held when exposed to the same climate and seasonal factors.

I performed pattern analysis of the birth month – disease risk curves for the circulatory system conditions between MSH and CUMC (set of 108 conditions). This was done to ascertain whether or not the birth month – disease relationship was the same between the two institutions. Seven of nine CUMC findings had significantly similar patterns at MSH (**Figure 7**, **Table 6**). In **Figure 7,** I show the seasonal risk patterns at birth obtained from MSH and CUMC for all nine circulatory system conditions. Four of these nine circulatory system conditions were also significant at the phenome-wide level at MSH. Two associations, congestive cardiac failure and mitral valve disorder, did not have statistically significant patterns.

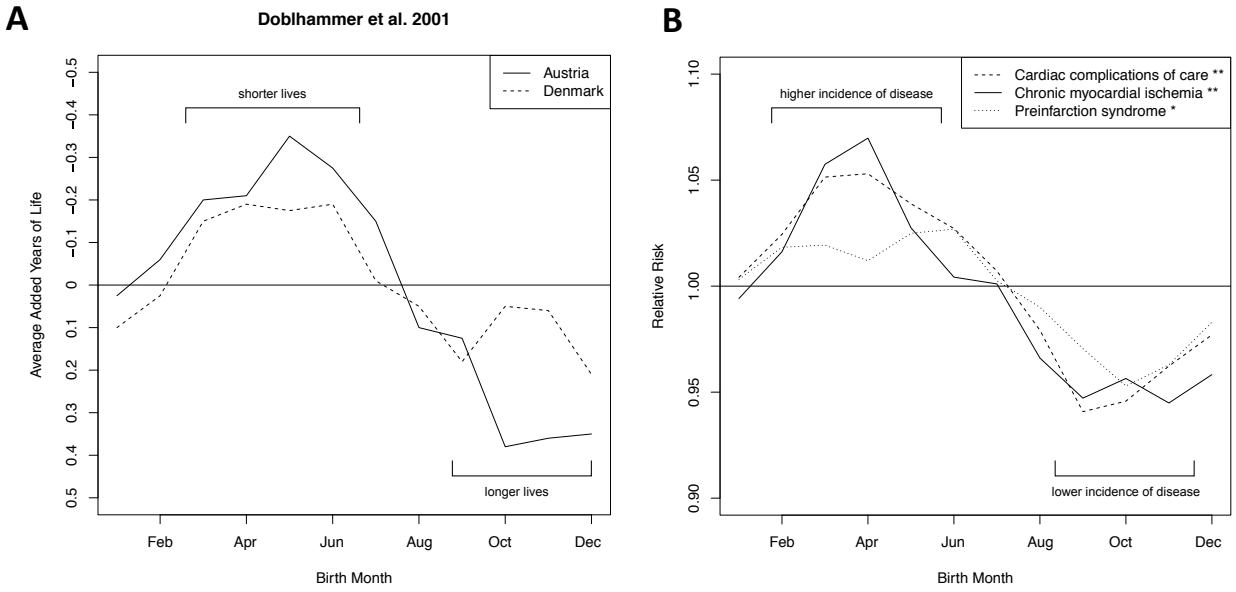


Figure 6. SeaWAS Cardiovascular Condition-Birth Month Proportions Correlate with Published Lifespan-Birth Month Results from Doblhammer et al. 2001. All 10 (9 novel) cardiovascular disease-birth month associations found by SeaWAS were compared to Doblhammer et al.'s lifespan-birth month dependencies for Denmark and Austria (Doblhammer and Vaupel, 2001). The lifespan-birth month associations are shown in **Figure 6a**. Six of the ten were anti-correlated (i.e., months with low cardiovascular disease risk were also months with longer life expectancies from Doblhammer et al.'s study (Doblhammer and Vaupel, 2001). The top three anti-correlated cardiovascular diseases are shown in **Figure 6b**, cardiac complications of care (Denmark: r=-0.815, p=0.001; Austria: r=-0.863, p<0.001); chronic myocardial ischemia (Denmark: r=-0.810, p=0.001; Austria: r=-0.826, p<0.001); and pre-infarction syndrome (Denmark: r=-0.712, p=0.009; Austria: r=-0.918, p<0.001). In **Figure 6b**, ** denotes P<=0.001 and * denotes P<0.01 for both comparisons (Austria and Denmark).

.

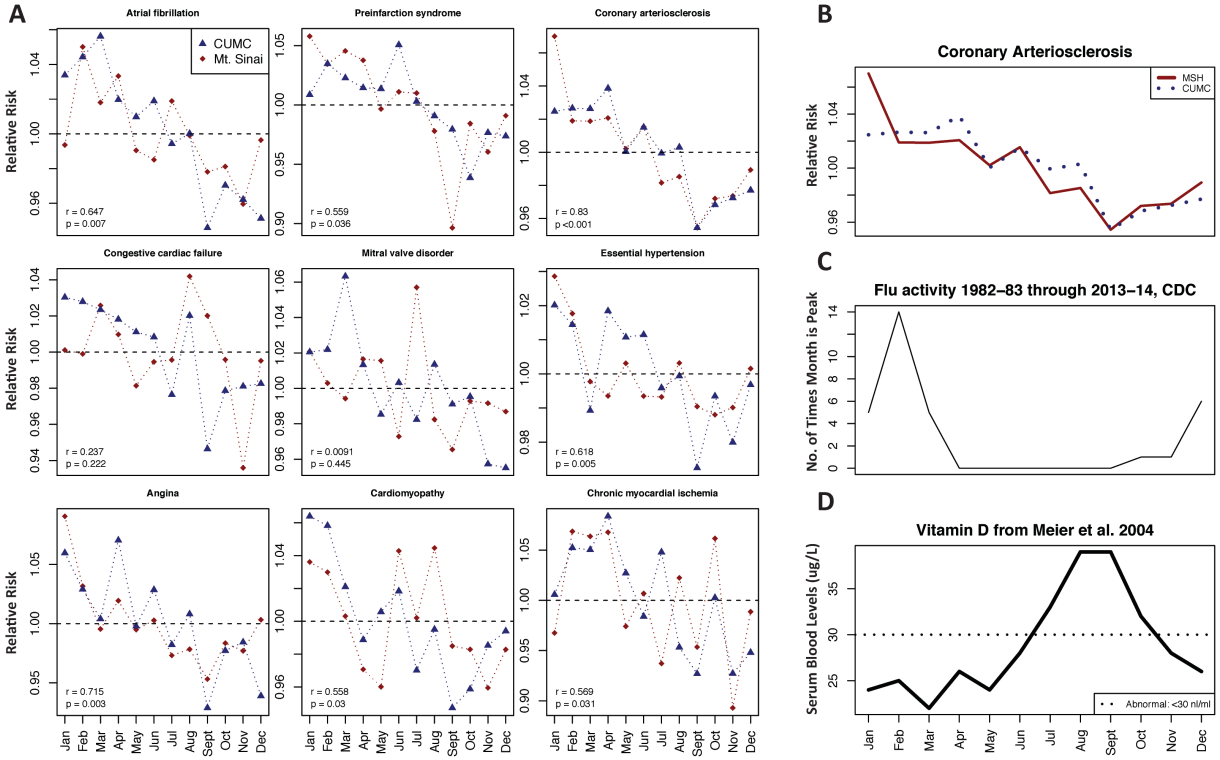


Figure 7. Cardiovascular Condition Risk vs. Birth Month Results from CUMC and MSH. **Figure 7A** shows results from all nine cardiovascular conditions from both MSH (red line) and CUMC (blue line). Seven of nine cardiovascular conditions were correlated at a statistically significant level with MSH data (i.e., the birth month – condition patterns were correlated) using Pearson's correlation. A significant pattern across the two institutions indicates that the birth month – condition relationship is the same. **Figure 7B** shows the most correlated result between MSH and CUMC was coronary arteriosclerosis ($r=0.83$, $p<0.001$). **Figure 7C** shows the comparison with the peak flu season month using CDC data on flu activity from 1982-83 through 2013-14 (URL: http://www.cdc.gov/flu/about/season/flu-season.htm). Serum vitamin D levels reported in Meier et al. 2004 (Meier et al., 2004) is also included in **Figure 7D**.

.

I overlaid data on seasonal variation in vitamin D levels and flu diagnosis levels to identify trends that could imply a biological rationale for observed cardiovascular – birth month association patterns. **Figure 7D** shows the seasonal variation in vitamin D levels (Meier et al., 2004) along with the birth month – coronary arteriosclerosis risk curves from MSH and CUMC. Low vitamin D months correspond to high coronary arteriosclerosis risk birth months (**Figure 7B**). In **Figure 7C,** I also included data from the CDC (http://www.cdc.gov/flu/about/season/flu-season.htm) containing the number of times each month was the peak month for a given flu season. This data was aggregated from 1982-83 through 2013-14 (CDC, 2014a).

2.4.10 Investigating Other Potential Confounders in Cardiovascular – Birth Month Findings

The relationship between birth month and cardiovascular disease was assessed in several stratified populations for 'Essential hypertension' – the most common cardiovascular disease (**Figure 8)**. Two common ethnicities – White and Black– were compared. I found that blacks had a slightly higher proportion of hypertension, but the overall birth month – hypertension curve was the same. For income status, I found the poor population to be at greater risk of hypertension overall then the rich population. The rich population was much smaller in size (a little over 11 thousand with hypertension), and the birth month – disease risk curve showed greatly variability. The overall trend for hypertension – birth month was similar across the sexes (female and male). In general, the trends for hypertension– birth month were the same across the stratified populations shown (**Figure 8**).

Table 6. Replication Results for Circulatory System Conditions Between MSH and CUMC: Phenome-Wide P-values and Pearson Correlation P-values

	Condition Type	Birth Month Risk - MSH		Birth Month Risk - CUMC		MSH	CUMC	MSH	CUMC	Pearson Corr.
Condition		**Low**	**High**	**Low**	**High**	**Max RR**	**Max RR**	**P***	**P ***	**P****
Atrial fibrillation	Symptom	11	2	9	3	1.050	1.056	0.226	$2.1X10^{-10}$	0.007
Coronary arteriosclerosis	Disease	9	1	9	4	1.070	1.039	$2.46X10^{-15}$	$3.1X10^{-8}$	<0.001
Essential hypertension	Symptom	10	1	9	1	1.029	1.020	0.003	$1.3X10^{-5}$	0.005
Congestive cardiac failure	State	11	8	9	1	1.042	1.030	0.760	$2.2X10^{-4}$	0.222
Angina	Symptom	9	1	9	4	1.091	1.070	0.002	$6.5X10^{-4}$	0.003
Cardiomyopathy	Disease	11	8	9	1	1.045	1.064	0.760	$8.6X10^{-3}$	0.030
Chronic myocardial ischemia	Event	11	2	9	4	1.069	1.084	0.834	0.022	0.031
Mitral valve disorder	Disease	9	7	12	3	1.057	1.063	0.562	0.024	0.445
Pre-infarction syndrome	Symptom	9	1	10	6	1.058	1.051	0.022	0.036	0.036

* Phenome-Wide P-value, FDR adjusted (1688 conditions for CUMC, 1433 for Mt. Sinai)

** Empirically Derived P-value Using Random Permutations of Mt Sinai's birth month distribution

Bold indicates phenome-wide significance was attained at both CUMC and Mt. Sinai

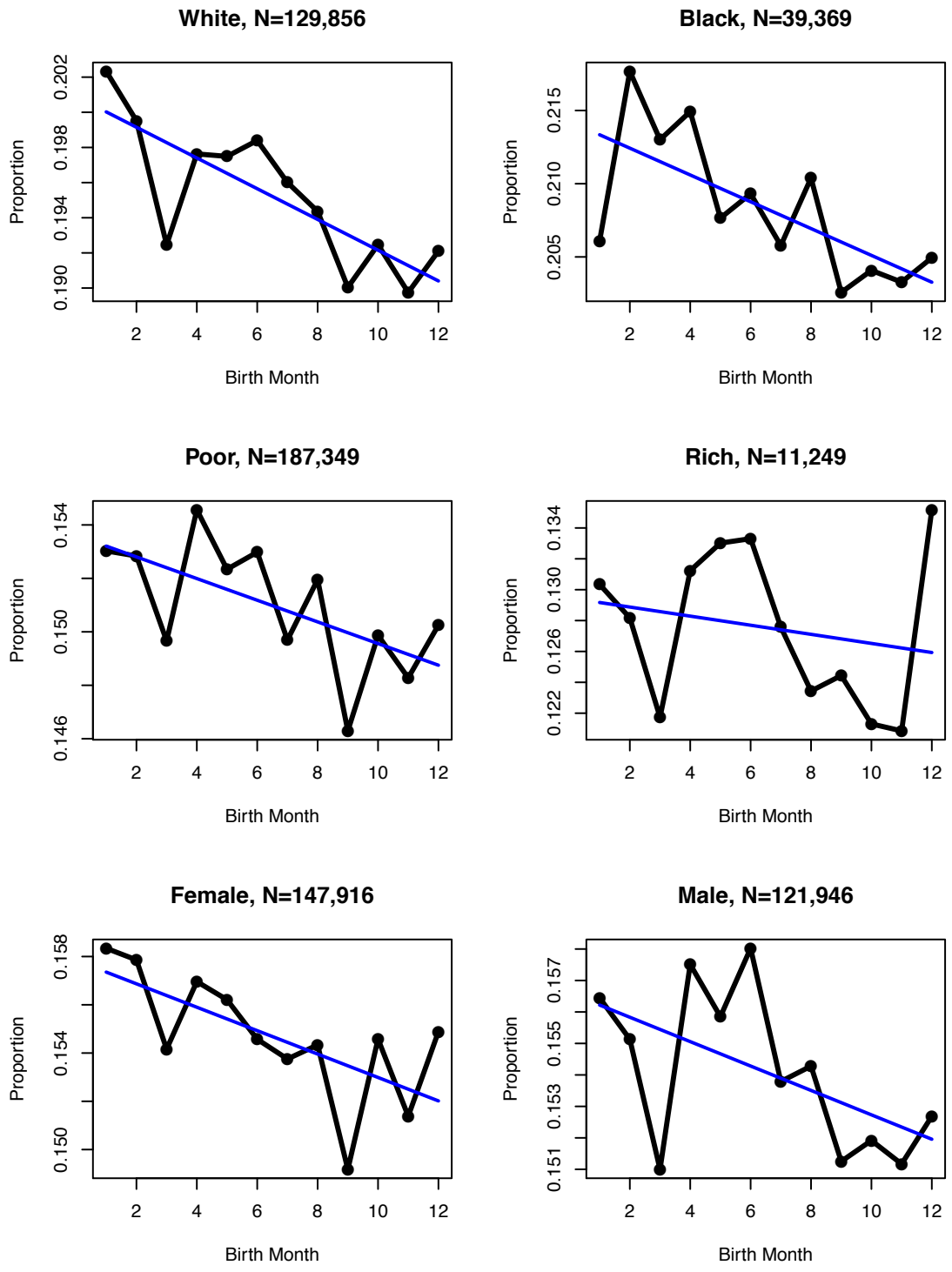


Figure 8. SeaWAS Hypertension-Birth Month Proportions By Ethnicity, Income and Sex. 'Essential hypertension' – birth month curves show the same trends across rich and poor and white and black populations. Males and females also showed similar trends. Larger differences were observed between rich and poor income groups then black and white ethnicities. However, the overall trends were the same indicating greater risk for hypertension in the winter months (Dec-Mar) and lower risk in the fall months (Sept-Nov). The blue line indicates the best-fit regression line for each curve.

2.5 Discussion

While the first 'SeaWAS' (SeaWAS: Season-Wide Association Study) study of its kind (Boland et al., 2015b), it was not the first study to investigate the role of climate on human health. The relationship between climate and climate factors on human health and disease is well known and studied (Dell et al., 2012; 2013; Epstein, 1999; Pöschl, 2005). Informatics researchers previously overlaid geospatial air quality data with human disease patterns to learn relationships between air quality and disease (Marinoni et al., 2015).

Many diseases demonstrate birth month dependencies with known mechanistic etiologies, including: asthma (Korsgaard and Dahl, 1983), ADHD (Halldner et al., 2014), reproductive performance (Huber et al., 2004), and myopia (Mandel et al., 2008). In these studies birth month was used as a proxy for seasonal variations in physiological state or changes in environmental exposures. Understanding dependencies between diseases and these variations is an important and challenging research task. My novel algorithm, SeaWAS uses a hypothesis-free method that does not relying on *a priori* hypotheses and makes use of large-scale population-level data obtained via EHRs.

2.5.1 SeaWAS Confirms Known Disease-Birth Month Associations

SeaWAS confirmed a literature-validated association between asthma (hyper-reactive airway disease) and birth month reported by studies from Denmark (Korsgaard and Dahl, 1983) and Sweden (Åberg, 1989). When I compared my findings to the Denmark study, I found a two-month shift in the birth month-asthma pattern that corresponds with a shift in the peak sunshine (a factor in asthma complicated by dust mite allergies) between Denmark and NYC (2014a; 2014b).

Likewise, comparing the reproductive performance results to an Austrian study (Huber et al., 2004) revealed that the dip in births among females born in May through September was observed in both studies (Huber et al., 2004). Importantly, the female reproductive system, unlike males, is established early with females being born with their lifetime maximum number of oocytes (Baker, 1963; Morita and Tilly, 1999). Oocyte count is thought to be linked to fertility (Tilly et al., 2009). Many studies show a link between maternal birth month and number of offspring supporting the belief that prenatal and early developmental effects can alter female's lifetime fertility (Huber et al., 2008; Huber and Fieder, 2009; 2011; Huber et al., 2004; Kemkes, 2010). SeaWAS findings bolster this body of literature.

I compared the ADHD results to those reported by a Swedish study and found a similar upward trend towards the later part of the year peaking in November (Halldner et al., 2014). The rationale for their findings that they reported was that relative age/immaturity (born later in the year) may result in increased ADHD detection (Halldner et al., 2014). This occurs because more immature children (i.e., younger in age) face higher demands early on in their school years making them more susceptible to ADHD diagnosis. The age cutoff for schools in Sweden is December 31st, which is the same for NYC public schools. Another alternative explanation is that the relationship between Vitamin D and ADHD and learning patterns has been established in rats (Becker et al., 2005; Burne et al., 2004) and Vitamin D deficiency in early development (*in utero* or shortly after birth) could be related to ADHD.

2.5.2 Discovered Cardiac Condition-Birth Month Relationship

SeaWAS revealed nine cardiovascular conditions associated with birth month. In the literature there are reports of a connection between perinatal flu exposure and increased risk of cardiovascular diseases later in life. Most notably, a study on children born to survivors of the

H1N1 1918 subtype who were associated with a >20% excess risk of cardiovascular disease (Mazumder et al., 2010). They also investigated the maternal malnutrition hypothesis and found that the maternal infection and cardiovascular disease risk relationship was independent of maternal malnutrition (Mazumder et al., 2010). Therefore, maternal infection during the winter months (Jan-Mar) could contribute to the increased cardiovascular disease risk among children born in those months that I observed in NYC.

Looking at all ten (nine novel) cardiovascular conditions revealed that individuals born in the autumn (September-December) were protected against cardiovascular conditions while those born in the winter (January-March) and spring (April-June) were associated with increased cardiovascular disease risk. Interestingly, a study found that people born in the autumn (October-December) lived longer than those born in the spring (April-June) (Doblhammer and Vaupel, 2001). Furthermore the relationship between cardiovascular disease risk and lifespan is established (Stamler et al., 1999). I compared the results to the Doblhammer *et al.* study investigating lifespan's dependency on birth month and found six cardiovascular diseases were significantly anti-correlated. This indicates that birth months with low risk for six cardiovascular diseases were also associated with longer lifespan in Doblhammer's study (Doblhammer and Vaupel, 2001) (**Figure 6)**. The findings suggest that the relationship between lifespan and birth month (Doblhammer and Vaupel, 2001) could possibly be explained by increased cardiovascular disease risk.

The relationship between cardiovascular disease and birth month could be mediated through a developmental Vitamin D-related pathway. Serum 25-hydroxyvitamin D levels are lower and parathyroid hormone levels are higher during the winter when no supplementation is given (Dawson-Hughes et al., 1991). Even with maternal supplementation, seasonally dependent

Vitamin D deficiency has been observed among breastfed infants (Halicioglu et al., 2012) and newborns (Lee et al., 2007). This is important because levels of parathyroid hormone and Vitamin D are associated with cardiovascular disease (Lee et al., 2008; Wang et al., 2008). Specifically elevated parathyroid hormone is correlated with increased heart failure in elderly males (Wannamethee et al., 2014). Studies focusing on adolescents found that Vitamin D deficiency resulted in an increased likelihood of hypertension (a SeaWAS discovered association) (Kumar et al., 2009; Reis et al., 2009) and high-density lipoprotein cholesterol (Kumar et al., 2009), both risk factors for cardiovascular disease.

2.5.3 Replicating Novel Cardiovascular Disease – Birth Month Relationships at Mount Sinai Hospital

Coronary arteriosclerosis, essential hypertension, angina and pre-infarction syndrome were all significantly associated with birth month at the phenome-wide level at both MSH (after adjusting for 1433 tests) and CUMC (after adjusting for 1688 tests). Their seasonal patterns were also significantly correlated (**Table 6, Figure 7**) supporting the connection between these conditions and birth month. While the patterns were highly correlated, in some instances the birth month with the max Relative Risk (RR) differed between the two institutions. For example, angina and coronary arteriosclerosis both experienced maximum RR's in April using CUMC data versus January for the MSH data. However, the patterns for both of these conditions were highly correlated (coronary arteriosclerosis, $r = 0.830$; angina, $r = 0.715$). This underscores the importance of analyzing the entire pattern (**Figure 7**) versus merely selecting maximum or minimum birth months to avoid misleading results.

Three conditions, atrial fibrillation, cardiomyopathy, and chronic myocardial ischemia, were significantly correlated between CUMC and MSH but were not significantly associated with

birth month at the phenome-wide level at MSH. These conditions could be associated with birth month through their common comorbidities, namely essential hypertension, angina, coronary arteriosclerosis and pre-infarction syndrome (Benjamin et al., 1998). However, it could also be the effect of sample size between the two institutions and the effects of stringent phenome-wide p-value adjustment. Many of these conditions had lower maximum RRs at MSH. For example, chronic myocardial ischemia had a maximum RR of 1.069 (0.973, 1.174) at MSH vs. 1.084 (1.011, 1.162) at CUMC. The significant correlation of the patterns between the two institutions supports the hypothesis of an underlying birth month effect for these conditions.

There are several additional reasons why results from CUMC and MSH may differ. Firstly, many of these circulatory system conditions are comorbid with one another (e.g., atrial fibrillation often occurs with other conditions such as essential hypertension) (Benjamin et al., 1998). Conditions that replicated across institutions were 'core' heart conditions, such as essential hypertension, that often occur in the presence of other cardiac ailments (Blair et al., 1996). These conditions may actually be driving the disease risk – birth month association. If this is the case, then they remain significant across institutions (provided the climate is the same). While both hospitals are located in NYC, they may serve patients with different socioeconomic demographics. These differences could play a role in the accessibility of certain healthcare options. Importantly, in spite of these issues, seven of nine CUMC circulatory system findings had statistically significant patterns at MSH.

2.5.4 Proposed Biological Mechanisms Underlying Cardiovascular Disease – Birth Month Relationships

Several mechanisms have been proposed previously that could explain the relationship between cardiovascular conditions and birth month, first described in (Boland et al., 2015b). Vitamin D

deficiency is known to increase cardiovascular disease risk, this is especially true for patients who already have essential hypertension (Lee et al., 2008). Furthermore, vitamin D levels have been shown to vary seasonally in women (Meier et al., 2004). Vitamin D levels in babies also depends on maternal vitamin D (Lee et al., 2007). Another hypothesis to explain cardiovascular – birth month relationships is maternal flu infection. Researchers found that children born to survivors of the H1N1 1918 subtype had a >20% excess risk of cardiovascular disease later in life (Mazumder et al., 2010). Maternal infection tends to be higher in the winter months (January – March) therefore this could contribute to increased risk among children born in those months.

Figure 7 shows that low vitamin D months correspond to high coronary arteriosclerosis risk birth months. However when vitamin D is low (**Figure 7D)**, the risk for flu infection is also high (**Figure 7C**). It is unclear which mechanism (maternal flu infection and vitamin D) is potentially responsible for the birth month association. It could also be a combination of both mechanisms. Flu infection is high during the birth months with the highest risk, but vitamin D levels are also at their lowest during this same period (**Figure 7C, 7D)**. Also vitamin D plays a role in immune response (Cantorna et al., 2004; Mora et al., 2008), which could indicate that peak flu season and low vitamin D occur together for a mechanistic reason (and are not independent of each other). My work confirmed a link between cardiovascular conditions and birth month across two institutions – MSH and CUMC – in the same climate (NYC) (Boland et al., 2015b). This supports the hypothesis that a biological mechanism tied to climate and seasonality could explain this increase in disease risk.

Birth months that were high in serum vitamin D (Jul-Oct.) appeared ideal for lower coronary arteriosclerosis risk. Additionally, birth months with a high flu burden (Jan – Mar.) were high-risk birth months for coronary arteriosclerosis. This does **not** indicate that being born in flu

season causes coronary arteriosclerosis later in life **nor** does it indicate that being born in a high vitamin D season lowers risk of coronary arteriosclerosis. These findings merely show support for proposed biological mechanisms, which require further validation from biologists, including more systematic examination in chapter 3 of this dissertation.

2.6 Limitations

Study limitations include the lack of condition independence (conditions rarely occur in isolation) potentially affecting multiplicity correction. Also, the algorithm cannot rule out indirect mechanisms (e.g., depression affects fertility, and learning ability) behind associations between disease risk and birth month. Some conditions associated with birth month may be associated because the infant was born in a high-risk period, e.g., acute bronchiolitis-autumn births. These associations differ from lifetime disease effects, however this algorithm does not distinguish between them because both are presented in the literature as birth month-disease associations (chapter three does address some of this effect). Another limitation is the exclusive use of EHR data, which is affected by the healthcare process (Hripcsak and Albers, 2013) and can introduce bias (Hripcsak et al., 2011), e.g., sick patients tend to be over-represented in EHR populations (Weiskopf et al., 2013). The potential affect of this bias is minimized in my findings because the birth month by year data at CUMC correlated with CDC data on individuals born in NYC counties. This indicates that CUMC's EHR population adequately represents the 'true' NYC-born population (which includes healthy people) with respect to birth month.

An issue with this version of the SeaWAS algorithm is that birth month is modeled as a numerical variable and not an ordinal variable. This can result in an over-estimation of the p-value for birth month – disease relationships with certain non-linear shapes. This would result in those shapes being determined as not dependent on birth month when in fact there is a non-linear

dependency. Therefore, there might be more than 55 conditions associated with birth month at NYC then those found by this initial SeaWAS algorithm. Importantly, another algorithm presented in chapter three of this dissertation, correlates specific exposures with birth month – disease risk dependencies curves. Therefore the algorithm in chapter three is not affected by this numerical modeling of birth month issue.

2.7 Conclusion

This work presents my high-throughput algorithm called SeaWAS that uncovers conditions associated with birth month without relying on *a priori* hypotheses. SeaWAS confirms many known connections between birth month and disease including: reproductive performance, ADHD, asthma, colitis, eye conditions, otitis media (ear infection) and respiratory syncytial virus. The algorithm revealed 16 associations with birth month that have never been explicitly studied previously. Nine of these associations were related to cardiovascular conditions strengthening the link between cardiac conditions, early development, and Vitamin D. Seasonally-dependent early developmental mechanisms might play a role in increasing lifetime disease risk.

2.8 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in the Journal of the American Medical Informatics Association (original SeaWAS algorithm) (Boland et al., 2015b) and Scientific Reports (Replication Study of Cardiovascular Findings) (Li et al., 2016). With a special thanks to Riccardo Miotto and Li Li from Mount Sinai Hospital who helped with refining the algorithm to suite their internal clinical data structure.

Chapter 3

Detection of Environmental Drivers at Birth that are Instrumental in Later Risk of Disease

3.1 Abstract

Birth month and climate are known to impact lifetime disease risk while the underlying exposures remain largely elusive. The purpose of this study is to uncover distal causal risk factors underlying birth month – disease relationships. This global birth month – disease study includes data from ten and a half million individuals from six sites, three countries, and four distinct climates to probe the relationship between exposure variance across sites and disease risk variance by birth season. My method couples Electronic Health Records (EHRs) with the seasonality of climate factors, pollutants, and influenza-like illness to uncover the most informative birth season exposures. Correlations are performed between each birth month - disease risk curve and a host of identified exposures at each study site. A meta-analysis of these correlations determines the most significant birth month – exposure relationships across all six-study sites. The algorithm adjusts for multiplicity using Bonferroni's method to select only the

most robust relationships. My method also successfully distinguishes relative age effects (a cultural effect) from environmental exposures. The only identified relative age association was Attention Deficit Hyperactivity Disorder. First trimester exposure to carbon monoxide was linked with increased risk of depressive disorder. I also found first trimester exposure to fine air particulates increased atrial fibrillation risk. Decreased exposure to sunlight during the third trimester increased the risk of type 2 diabetes mellitus. Global study of birth month – disease relationships reveals distal causal risk factors underlying these relationships. Important biological hypotheses can be formed only when the distal factors have been identified.

3.2 Introduction

Seasonality and climate together play an important role in human health and disease, which has been studied for millennia (Hippocrates and Galen, 1952). Geographic location alone alters the exposure to many diverse environmental exposures (Jarup, 2004). Variance in these exposures is important in understanding differences in disease risk among populations. Prenatal or perinatal exposure to many environmental variables has been tied with increased disease risk later in life. This includes climate factors such as reduced sunlight (Waldie et al., 2000) and high humidity (Crowther, 1985). Flu, or influenza-like illness (ILI) exposure during pregnancy is also tied to increased disease risk in offspring (Bánhidy et al., 2005). Furthermore, exposure to pollutants during pregnancy can increase risk of disease among the offspring. Such pollutants include carbon monoxide (Raub et al., 2000), nitrogen dioxide (Maroziene and Grazuleviciene, 2002; Singh, 1987), ozone (Salam et al., 2005), and sulfur dioxide (Rogers et al., 2000). These exposures are also known to vary seasonally because of changes in atmospheric boundary layer depth, changes in emission rates and changes in wind and advection. Therefore it is reasonable to

hypothesize that seasonal variations in these factors could modulate birth month – disease risk patterns observed in epidemiology studies.

Electronic Health Records (EHRs) are currently used throughout the world to record and store health information collected during the clinical encounter. These EHRs represent a rich data source for high-throughput explorations of birth season – outcome relationships. A comprehensive analysis should shed light on the exposures driving the underlying birth season – disease effects.

My algorithm, described in chapter two, called SeaWAS for Season-Wide Association Study, systematically investigates birth month - disease dependencies across all diseases having sufficient prevalence in EHRs where birth month serves as a proxy for seasonal variance at birth. Initial studies were conducted using data from New York City (NYC). Novel cardiovascular findings were validated in a separate EHR system with increased disease risk being observed in winter months (Jan-April) (Li et al., 2016). This EHR was also in NYC and thereby subjected to similar climate constraints (Li et al., 2016). Previous studies did not identify environmental factors behind the associations because they were conducted in a single climate. Separately, researchers from North Russia (Northern Kola Peninsula) found that males born in the summer – fall had increased elasticity of blood vessels, which could be protective against cardiovascular disease later in life. Additionally, their results point to differences in cardiovascular physiology that appear birth month/season dependent (Melnikov et al., 2016). They also found that females who died from acute myocardial infarction (AMI) were found to have a significant birth season relationship in the Sakha Republic, Russia (Melnikov, 2003).

In this study, I develop another algorithm to investigate the relationship between developmental stages (first, second, third trimester, perinatal or pregnancy-wide) and seasonal environmental

exposures (climate, pollution, flu) for birth month – disease relationships. The algorithm delineates birth month – disease relationships due to differences in school cutoff dates across sites indicating the effect of relative age on human health and disease. I present results obtained using data from six distinct institutions, over three countries, spanning five cities, and four distinct climates.

3.3 Methods

3.3.1 Clinical Data

Birth month – disease risk data were obtained from six different hospitals or study sites. A version of the SeaWAS algorithm was published on GitHub (Boland, 2015) that conforms to the Common Data Model (CDM) adopted by the Observational Health Data Sciences and Informatics (OHDSI) consortium (Overhage et al., 2012) allowing my code to be shared among the OHDSI community. A SeaWAS was performed at three OHDSI collaborator sites using OHDSI-formatted R scripts that were run locally on each site's EHR databases. Three study sites were not OHDSI participants at the time of the study. Therefore code was formatted to meet their individual institution's data schemas. Permission was obtained from each institution's local Institutional Review Board (IRB), which conforms to each country's, and in some cases state's, laws and guidelines.

A call for members of the Observational Health Data Sciences and Informatics (OHDSI) group was put together to solicit collaborators from within the OHDSI community that already have data conformed to the OMOP CDM (Boland et al., 2015a). The call for collaborators began in September 2015. Initially, 15 sites (not including CUMC) agreed to participate in some way including 6 non-OHDSI member sites (total: 9 OHDSI collaborator sites, 6 non-OHDSI collaborator sites). **Table 7** shows the success-in-obtaining-data rates for OHDSI vs. non-OHDSI

collaborators when running SeaWAS. Typically, higher rates of engagement were seen among non-OHDSI members because these individuals were excited and personally motivated to participate in the research. Soliciting interest within the OHDSI community was a challenging issue and many hurdles were encountered due to differences among various institutions' methods for handling IRBs. This represents one of the challenges of collaborating on a global scale (Boland et al., 2017a). Issues that resulted in studies dropping out completely were usually centered on institution-specific data privacy restrictions that were insurmountable. For example, some institutions only permitted de-identified data to be mapped to the OMOP CDM. Birth month is a protected health data element meaning that any database containing birth month is considered to be potentially re-identifiable. This seemingly nuanced legal distinction resulted in the final exclusion of several sites.

Table 7. Success and Failure Rates in Obtaining Data for the Extended SeaWAS Study

Status of Data	OHDSI member?	% Rate
Data Submitted to Study – Success	3 OHDSI sites	30.00% OHDSI
	3 non-OHDSI collaborator site	50.00% non-OHDSI
Stuck in IRB/Other Issue	4 OHDSI sites	40.00% OHDSI
	3 non-OHDSI collaborators	50.00% non-OHDSI
Dropped out of Study Completely Before IRB	3 OHDSI sites	30.00% OHDSI
Total	**16 (10 OHDSI, 6 non-OHDSI)**	**37.50% Success Rate (6/16)**

The algorithm maps International Classification of Diseases, version 9 (ICD-9) codes to the Systemized Nomenclature for Medicine – Clinical Terms (SNOMED-CT) codes using the Common Data Model (Boland et al., 2015b; Overhage et al., 2012). Only the first instance of a diagnosis for each patient was included in the algorithm (for full algorithm details see Boland *et*

al.) (Boland et al., 2015b). Hereafter, I refer to distinct medical SNOMED-CT diagnoses as diseases realizing that some may be indicative of medical conditions.

First, site characteristics were obtained, including, patient demographics, setting (climate, in-patient/out-patient) and CDM version number (if an OHDSI data partner). For climate, the Köppen-Geiger climate classification system (Köppen, 1884; Kottek et al., 2006) was used to describe the high-level climate of each region.

The SeaWAS algorithm returns birth month – disease risk curves for all diseases with at least 1,000 patients at a given site. These disease risk – birth month curves were then used as input into the developmental time point – exposure – disease model described below in the Statistical Modeling section.

3.3.2 Exposure Data

To study the relationship between exposure and birth season relationships, a dataset was required containing seasonal variance in exposures across a variety of exposure types and locations. I investigate 6 climate variables (mean sunshine hours, minimum temperature, maximum temperature, rainfall in inches, relative humidity, days of precipitation), 5 pollutant variables (fine particulate matter (PM 2.5 micrometers in diameter), ozone (O3), carbon monoxide (CO), nitrogen dioxide (NO2), and sulfur dioxide (SO2)), and flu / influenza-like illness (ILI) in this study. **Figure 9** illustrates the variation in seasonal exposure to each of the 12 factors (climate, pollution and influenza) across all sites. Exposure data were assembled from the Centers for Disease Prevention and Control (CDC), Environmental Protection Agency (EPA) and the National Oceanic and Atmospheric Administration (NOAA). For Taiwanese and Korean data, I obtained data from the Korean Meteorological Administration, the Taiwanese Central Weather Bureau, and the Korean CDC (KCDC) Virological Surveillance data. When data was unavailable

in a freely accessible public dataset, I used published literature to obtain the required seasonality in pollutant or flu exposure information.

3.3.3 Statistical Modeling

Because cultural/sociological effects of birth month are fundamentally different from environmental effects, I first determined those diseases with significant relative age effects (Musch and Grondin, 2001). Relative age was defined as an individual's age relative to the individual's peers in the same school grade. Next, I investigated the relationship between seasonal environmental exposures and birth month – disease risk.

3.3.4 Delineating Culture Effects from Seasonal Environmental Effects

The first step in modeling the relationship between birth month – disease risk and various exposures, was to distinguish birth month effects that were driven by purely cultural elements from those due to exposure to environment, pollution or some other factor. For instance, in sports, the age of each child athlete relative to their peers determines their ability to succeed. This has been demonstrated in multiple cases (Helsen et al., 2005; Musch and Grondin, 2001) and has been characterized as the 'relative age effect'. Children that are 'older' relative to their peers are more likely to succeed in athletics, whereas children 'younger' then their peers are at increased risk of being victims of bullying (Mühlenweg, 2010). To study the 'relative age effect', the public school cutoff dates were collected for each study site. Data were adjusted from each institution using the cutoff dates from that region. Therefore, curves ranged from 6 months older than the average child (i.e., just *after* the cutoff date) versus 6 months younger than the average child (i.e., just *before* the cutoff date).

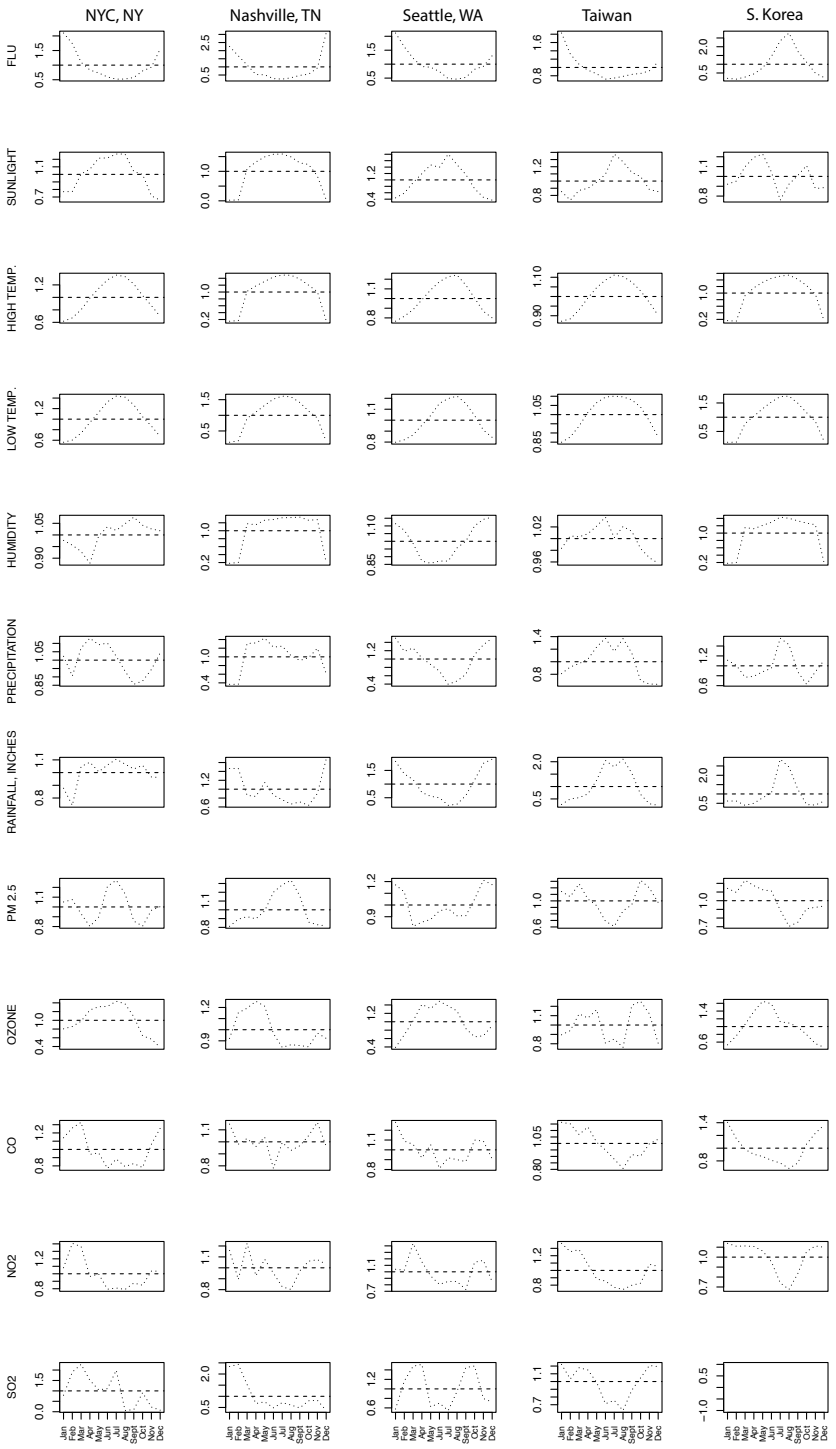


Figure 9. Seasonal Variance in Exposure to Twelve Different Factors. Six climate factors are included: mean sunshine hours, low temperature, high temperature, rainfall in inches, relative humidity, days of precipitation. Five pollutant factors included in this study: fine particulate matter (PM 2.5 micrometers in diameter), ozone (O3), carbon monoxide (CO), nitrogen dioxide (NO2), and sulfur dioxide (SO2) and Influenza-like illness (ILI).

A regression model for the relationship between relative age (+6 months vs. average…-6 months vs. average) and disease risk was used to compute the significance of relative age for each disease at each site. Diseases that were nominally significant across all 6 sites were considered to have significant cultural effects.

3.3.5 Modeling Seasonal Environmental Exposures Occurring During Developmental

Twelve seasonally varying environmental exposures were identified as potential factors involved in birth month – disease relationships (**Figure 10).** To model the relationship between exposures and birth month – disease risk, I first modeled the exposure level for each critical developmental time point. The trimester of an exposure is vital in determining the effects on the offspring (Bérard et al., 2007; Goldstein, 1995), therefore I examined the cumulative exposure for each factor across each of the three trimesters. In addition, I investigated pregnancy-wide exposure (cumulative exposure across the entire pregnancy) and perinatal exposure (exposure at birth) as these also represent critical developmental periods.

The average gestation period in weeks was obtained for each country: USA, South Korea and Taiwan. The mean gestation was 38.5 weeks in Taiwan (Hsieh et al., 2006), 39.17 weeks in South Korea (Lim et al., 2010) and 38.6 weeks in the USA according to the CDC (CDC, 2015). These average gestation periods were used to compute the typical conception month for each birth month.

Next the cumulative exposure for each developmental stage (e.g., first trimester) for each factor (e.g., sunlight, rainfall) was calculated for a given birth month. These calculations were made using the midpoints of each month. For example, an October birth month would have a typical first trimester period from mid-January – mid-April; a typical second trimester period of mid-April – mid-July; and a typical third trimester period of mid-July – mid-October. Therefore first

trimester sunlight exposure for an October birth month would include the sunlight exposure from mid-January through mid-April and so on.

3.3.6 Meta-Analysis Across All 6 Sites Using Random Effects Modeling

First, I correlated each exposure - developmental stage (e.g., first, second, third trimester) with the disease relative risk by birth month per site. Each disease was compared against each developmental time point for each factor (e.g., sunlight, rainfall, etc.). Pearson's correlation was determined for the relationship between the exposure during a certain period (e.g., first trimester) and disease risk. Because the seasonality of exposures varied widely across sites, these correlations were performed for each study site.

Next, I employed a meta-analysis approach to harness all data from the diverse sites. I used the DerSimonian-Laird (DSL) random-effect meta-analytical approach (DerSimonian and Laird, 1986) to determine an overall 'site-wide' correlation coefficient representing the effect of an individual exposure (e.g., sunlight) on a given disease (e.g., depression) during a specific developmental stage (e.g., first trimester). The DSL method transforms each site-specific correlation coefficient to a Fisher Z value with a standard error determined by the site-specific sample size. This weighs correlations from sites with a larger sample size for a given disease higher than correlations from sites with lower sample sizes. A summary correlation coefficient can then be computed from these 'sample-size adjusted' correlations. This summary statistic represents the overall correlation obtained from the meta-analysis across the 6 sites. The DSL method was implemented based on Schulze (Schulze, 2004) and incorporated in the R 'metacor' library (Laliberté and Laliberté, 2015) with widespread use among the research community (Polderman et al., 2015).

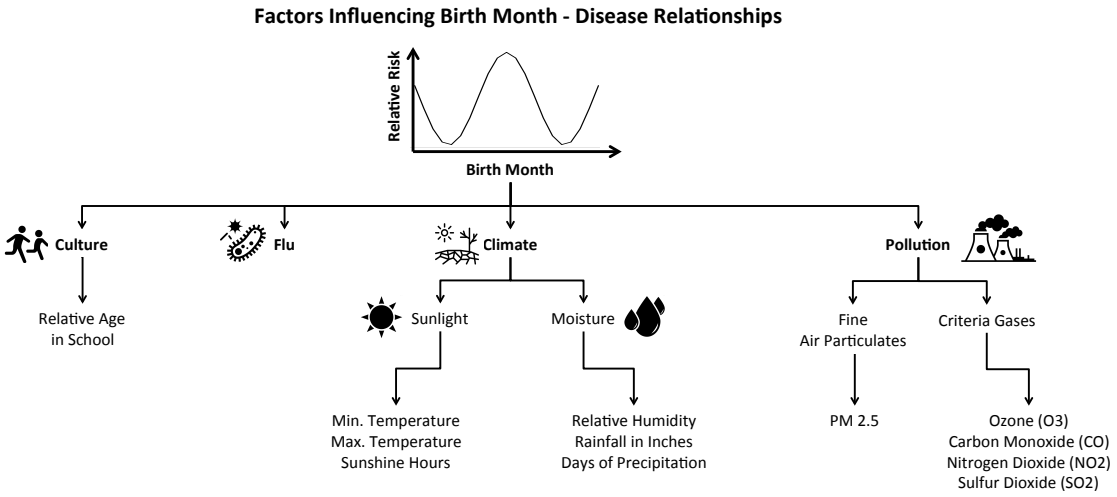


Figure 10. Factors that Could Influence Birth Month – Disease Relationships. This includes the cultural effect of relative age and 12 'traditional' exposures including flu / influenza-like illness, 6 climate variables (3 sunlight and 3 moisture-related), fine air particulates and 4 criteria gases including ozone, carbon monoxide, nitrogen dioxide and sulfur dioxide.

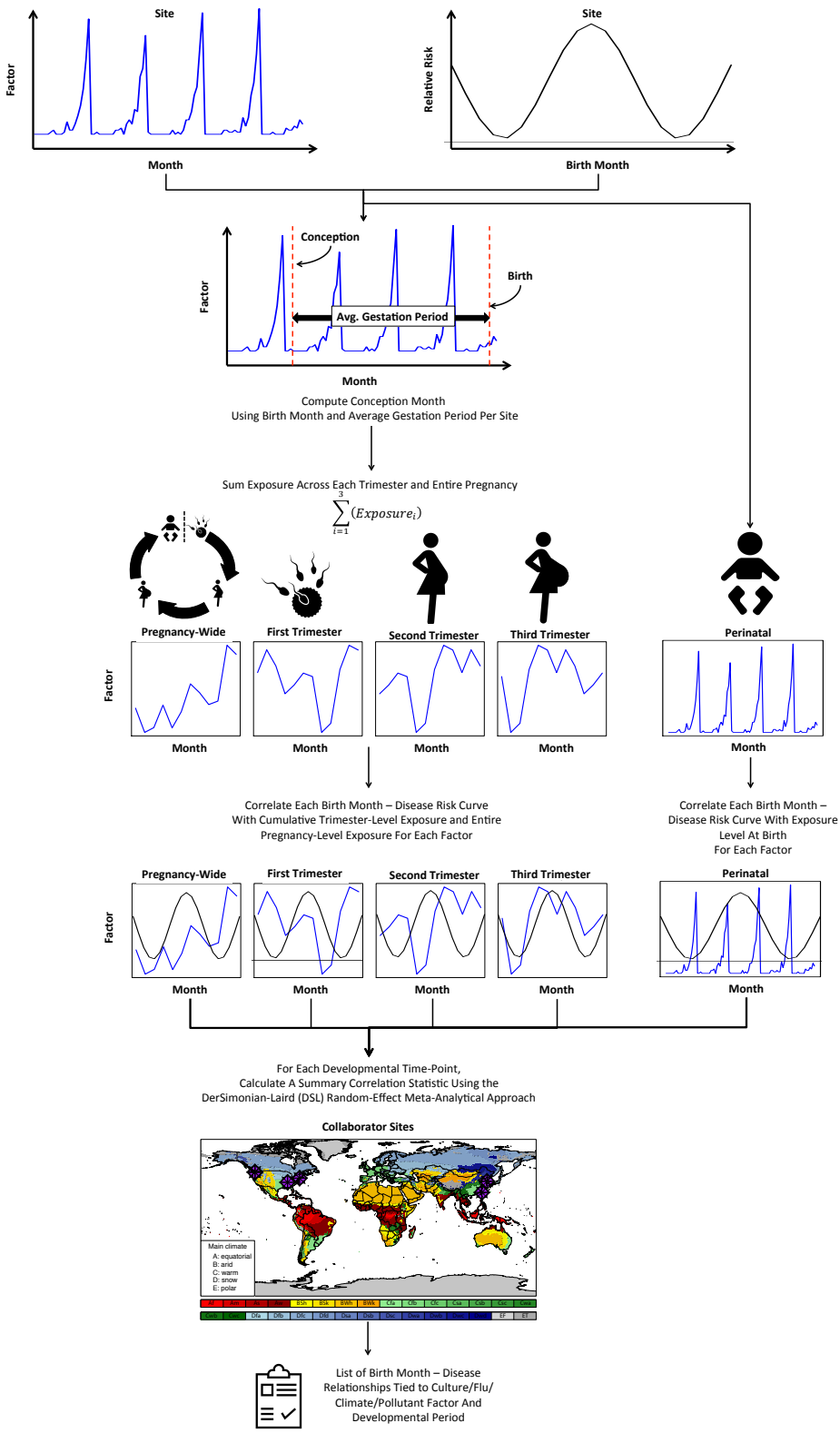


Figure 11. Schema Depicting the Model that Captures the Environmental Exposures' Effect At Various Developmental Time Points During Prenatal / Perinatal Development. Results are integrated across multiple sites using the DerSimonian-Laird Random Effects Meta-Analytical Approach.

Hence, my method determines the correlation between each of 12 exposures across 133 diseases during 5 different developmental stages (i.e., three trimesters, pregnancy-wide and perinatal). Therefore, multiple comparisons must be accounted for in the analysis. To remain as stringent as possible, and bias ourselves against finding disease – exposure relationships, I used Bonferroni's method of p-value correction that adjusts for all comparisons, including all 133 diseases, 12 exposures and 5 developmental stages (133x12x5=7980 tests). This stringent threshold allows me to state that ***exposure X*** during ***stage Y*** is associated with increased or decreased ***risk of disease Z***. **Figure 11** illustrates the overall method to find significant exposure-disease relationships for a given developmental stage.

3.4 Results

3.4.1 Data

Data were obtained from six study sites including Columbia University (CU) in New York City, New York; Mount Sinai Hospital (MSH) in New York City, New York; Vanderbilt University (VU) in Nashville Tennessee; University of Washington (UW) in Seattle, Washington; Ajou University in Suwon, South Korea and the Taiwan National Health Insurance program, which contains data from each of Taiwan's four geographic regions. **Table 8** contains a breakdown of the patient demographics from each study site. Overall, patients were middle aged ranging from a median of 35 years old in Taiwan to 53 years old at MSH in NYC. However, most datasets had a median age in the 40s. Race and ethnicity varied by site due to differences in local populations. Both datasets from Asia: Taiwan and South Korea did not collect race/ethnicity data and therefore only nationality was used with the assumption that the majority of patients were Asian. The percent of Hispanic patients also varied across sites with 2-4% at UW and VU versus 17-21% Hispanic at both NYC sites.

Table 8. Demographics of Patients Included in Climate-Wide SeaWAS (N=10,499,887)

Demographic	Columbia University N (%)	Mt Sinai N (%)	Vanderbilt University N (%)	University of Washington N (%)	Taipei Medical University N (%)	Ajou University School of Medicine N (%)
Location	New York City, New York	New York City, New York	Nashville, Tennessee	Seattle, Washington	Taiwan (99.99% of total population in Taiwan)	Suwon, South Korea
No. of Patients	1,749,400	1,169,599	3,051,997	1,770,510	909,689	1,848,692
Sex [1]						
Female	956,465 (54.67%)	678,717 (58.03%)	1,558,550 (51.07%)	895,351 (50.57%)	464,576 (51.07%)	892,178 (48.26%)
Male	791,534 (45.25%)	490,600 (41.95%)	1,278,939 (41.90%)	874,618 (49.40%)	445,113 (48.93%)	956,514 (51.74%)
Other [1]	1,401 (0.08%)	282 (0.02%)	214,508 (7.03%)	541 (0.03%)	-	-
Race						
White	665,366 (38.03%)	424,803 (36.32%)	1,653,093 (54.16%)	990,209 (55.93%)	NA	NA
Other [2]	456,185 (26.08%)	165,423 (14.14%)	NA	82,656 (4.67%)	NA	NA
Unidentified	386, 533 (22.10%)	256,819 (21.96%)	1,123,369 (36.81%)	367,100 (20.73%)	NA	NA
Black	189,123 (10.81%)	166,950 (14.27%)	241,978 (7.93%)	110,007 (6.21%)	NA	NA
Declined	29,747 (1.70%)	NA	5,638 (0.18%)	16,976 (0.96%)	NA	NA
Asian	20,746 (1.19%)	45,596 (3.90%)	24,109 (0.79%)	122,839 (6.94%)	NA	NA
Native American	1,511 (0.09%)	2,447 (0.21%)	3,074 (0.1%)	16,408 (0.93%)	NA	NA
Pacific Islander	189 (0.01%)	1,094 (0.09%)	736 (0.02%)	3,085 (0.17%)	NA	NA
Hispanic	(See Ethnicity)	106,467 (9.10%)	(See Ethnicity)	61,230 (3.46%)	NA	NA
Korean	NA	NA	NA	NA	NA	1,848,692 (100%)
Taiwanese	NA	NA	NA	NA	909,689 (100%)	NA
Ethnicity						
Non-Hispanic	590,386 (33.75%)	761,535 (65.11%)	713,853 (23.39%)	NA	NA	NA

Unidentified	458,071 (26.18%)	208,899 (17.86%)	2,280,039 (74.71%)	NA	NA	NA
Hispanic	361,123 (20.64%)	199,165 (17.03%)	44,527 (1.46%)	NA	NA	NA
Declined	339,820 (19.42%)	NA	13,578 (0.44%)	NA	NA	NA
Other Attributes		**Median (IQR)**				
Total SNOMED-CT codes per patient	6 (1, 32)	7 (3, 22)	8 (3, 26)	9 (3, 24)	186 (98, 338)	4 (2, 12)
Distinct SNOMED-CT codes per patient	3 (1,8)	5 (2, 10)	5 (2, 14)	4 (2, 11)	49 (33, 70)	4 (2, 12)
Age (year of service – year of birth)	38 (22, 58)	53* (36, 66)	44 (25, 61)	48 (34, 64)	35 (20, 50)	42 (28, 57)
Treatment Year Range	1985-2013	1979 - 2015	1991-2016	1993 - 2016	1998-2011	1994 - 2013
Köppen-Geiger Climate	Cfa	Cfa	Cfa	Csb	Aw	Dwa
In- / Out- Patient	In-patient	Both	Both	Both	Both	Both
CDM** Version	V.4	None	None	None	V.5	V.4

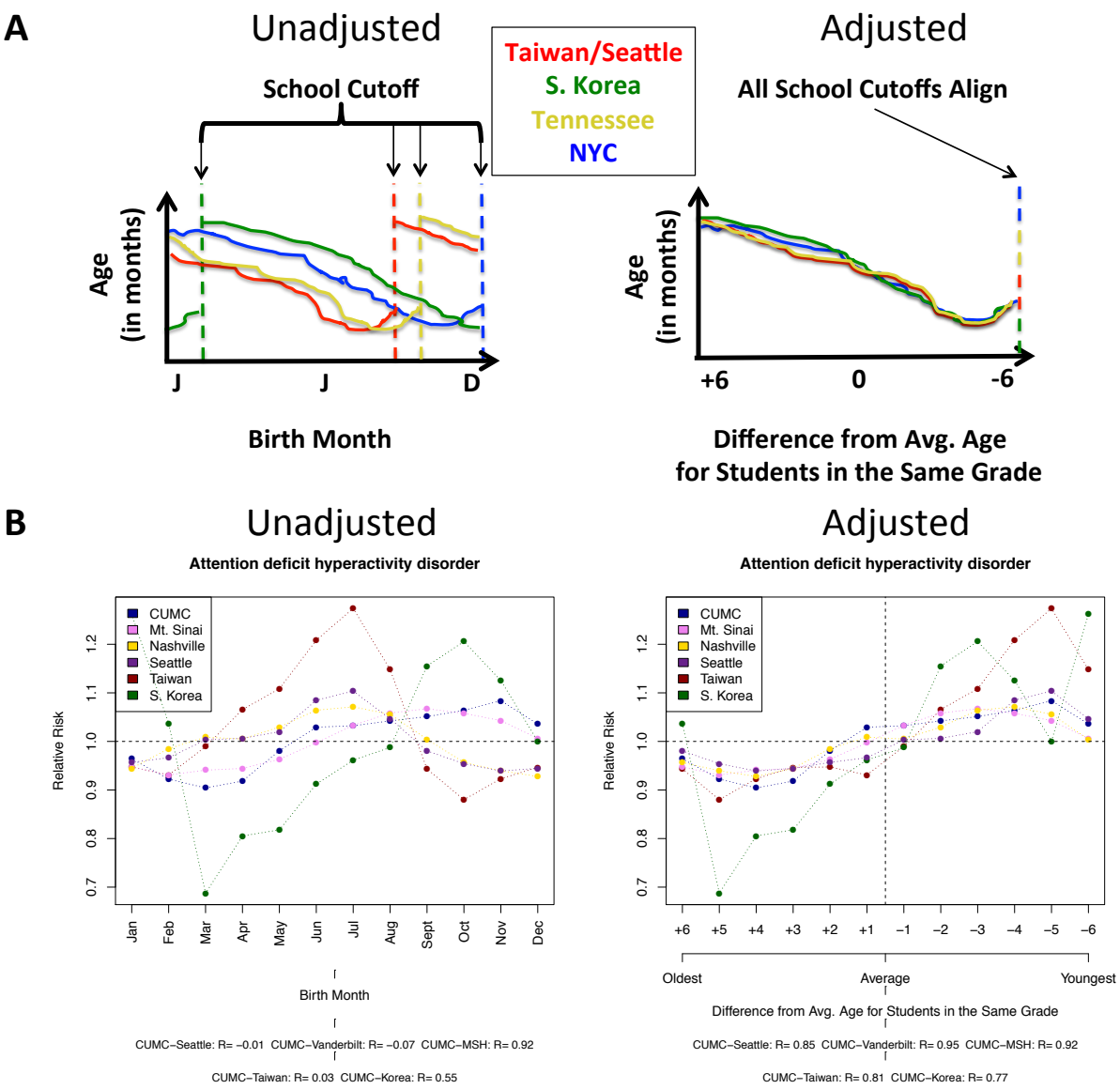
[1] Other (includes individuals of unidentified gender); [2] Other (includes Hispanics not otherwise identified); * Computed in days, age in years = age in days / 365.25; ** CDM: Common Data Model

I obtained the birth month – disease risk curves for each disease with at least 1000 patients at each study site. In total, 133 diseases had at least 1000 patients at all 6 sites and I focused on this intersection set for the remainder of the analyses. Disease-specific sample sizes varied across sites. 'Essential hypertension' was the most common disease at all four USA sites. Both Asian sites showed increased prevalence of gastrointestinal issues and lower incidence of cardiovascular disease.

3.4.2 Statistical Modeling

I first investigated the relationship between relative age –as determined by school cutoff dates– and the birth month – disease risk relationship. Out of 133 diseases, only one disease was significantly associated with relative age across all 6 sites – Attention Deficit Hyperactivity Disorder (ADHD). The result both before relative age adjustment (i.e., unadjusted birth month) and after is shown in **Figure 12.** The average difference in ADHD risk due to relative age was 17.97% (average peak of 1.084 vs. average trough of 0.904) with children younger then their peers experiencing greater ADHD risk. No other diseases were significantly correlated with relative age.

Next, the relationship between exposures at certain developmental stages (e.g., a given trimester) and disease risk were investigated. My method determines the correlation between each of 12 exposures across 133 diseases at five different developmental stages. Therefore, multiple comparisons must be accounted for in the analysis. **Figure 13** shows the Manhattan plot for each developmental stage. Results are reported as significant if they pass the Bonferroni correction threshold for multiple comparisons across all analyses (i.e., 133 diseases *12 exposures* 5 time points = 7980 tests).



NYC, CUMC P	NYC, MSH P	Nashville, TN P	Seattle, WA P	Taiwan P	S. Korea P
<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Figure 12. Method to Detect the Existence of a Relative Age Effect In Birth Month – Disease Associations and Results. Figure 1A illustrates the method of adjusting birth month – disease associations by school cutoff dates to calculate the relationship between relative age and disease risk. Taiwan and Seattle, Washington are grouped together because the school cutoff date is the same at both locations (Aug. 31). Figure 1B shows the only significantly associated disease found across all 6 sites between relative age and disease risk – Attention Deficit Hyperactivity Disorder (ADHD). The average difference in Relative Risk (RR) by relative age was calculated resulting in a difference of 17.97% in peak vs. trough months. Peak risk was observed in the -5 month and trough (lowest risk) was observed in the +4 month. Average peer age occurring at 0.

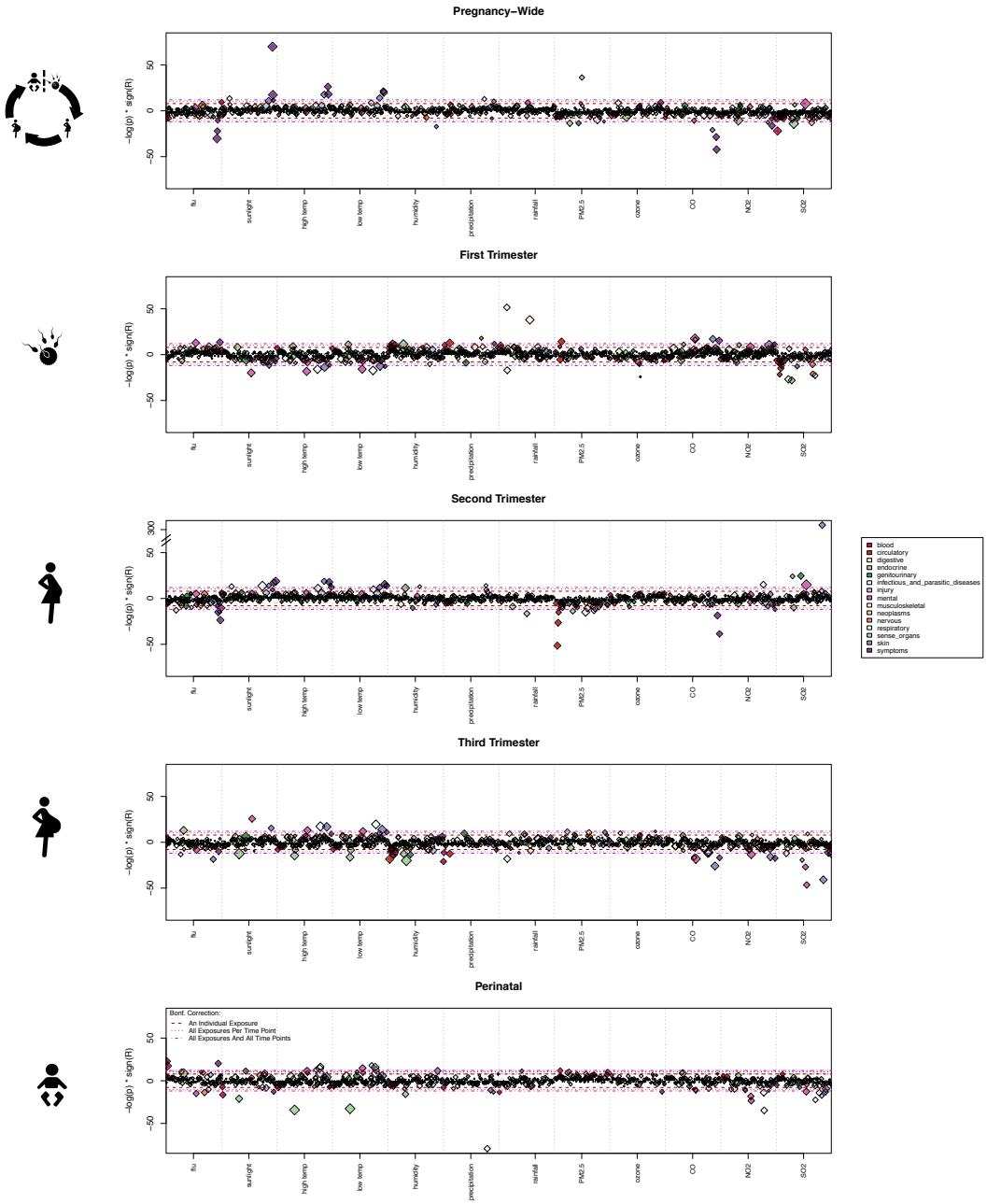


Figure 13. Manhattan Plot Showing Relationship Between Disease Risk and Exposures Occurring During Certain Developmental Time Points. Individual diseases are colored by their respective ICD-9 disease categories. The different Bonferroni adjusted demarcations are noted. Note that Acne is extremely associated with second trimester sulfur dioxide exposure (-log(p)>300). I reported results as significant if they pass the most stringent Bonferroni correction threshold (133 diseases *12 exposures* 5 time points = 7980 tests).

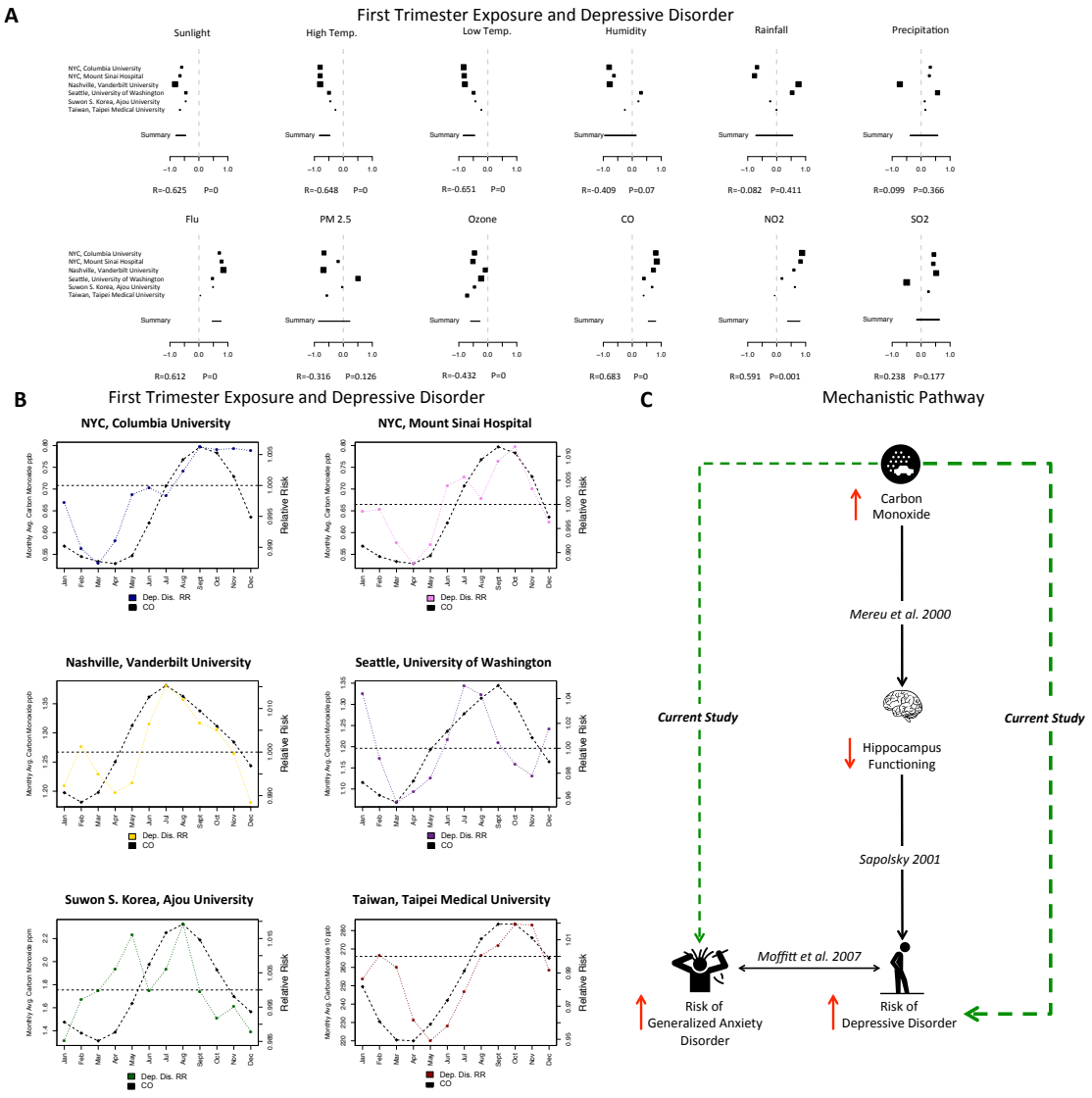


Figure 14. Depressive Disorder and First Trimester Exposure to Carbon Monoxide Figure 14A Depressive Disorder and First Trimester Exposure to All Environmental Factors. Larger squares in **Figure 14A** indicate correlations with larger confidence intervals, which typically occurs when the number of patients at a given site is low for a particular disease. **Figure 14B. Relationship Between Depressive Disorder and First Trimester Carbon Monoxide Exposure At Each Study Site.** Each site has its own subplot in **Figure 14B,** the colored line is the relative risk of depressive disorder at that site by birth month. The solid black lines indicate the first trimester exposure to Carbon Monoxide at each site. **Figure 14C. Connecting the Literature on First Trimester Carbon Monoxide (CO) Exposure and Offspring's Risk of Depressive Disorder and Our Current Study.** A solid black arrow denotes each literature link with directionality being denoted by up or down red arrows. The major link in my study is the link between first trimester CO exposure and increased risk of depressive disorder (thick dashed green line). I also found a lower correlation between GAD and first trimester exposure to CO suggesting that it could be patients afflicted with both diseases that were exposed to CO.

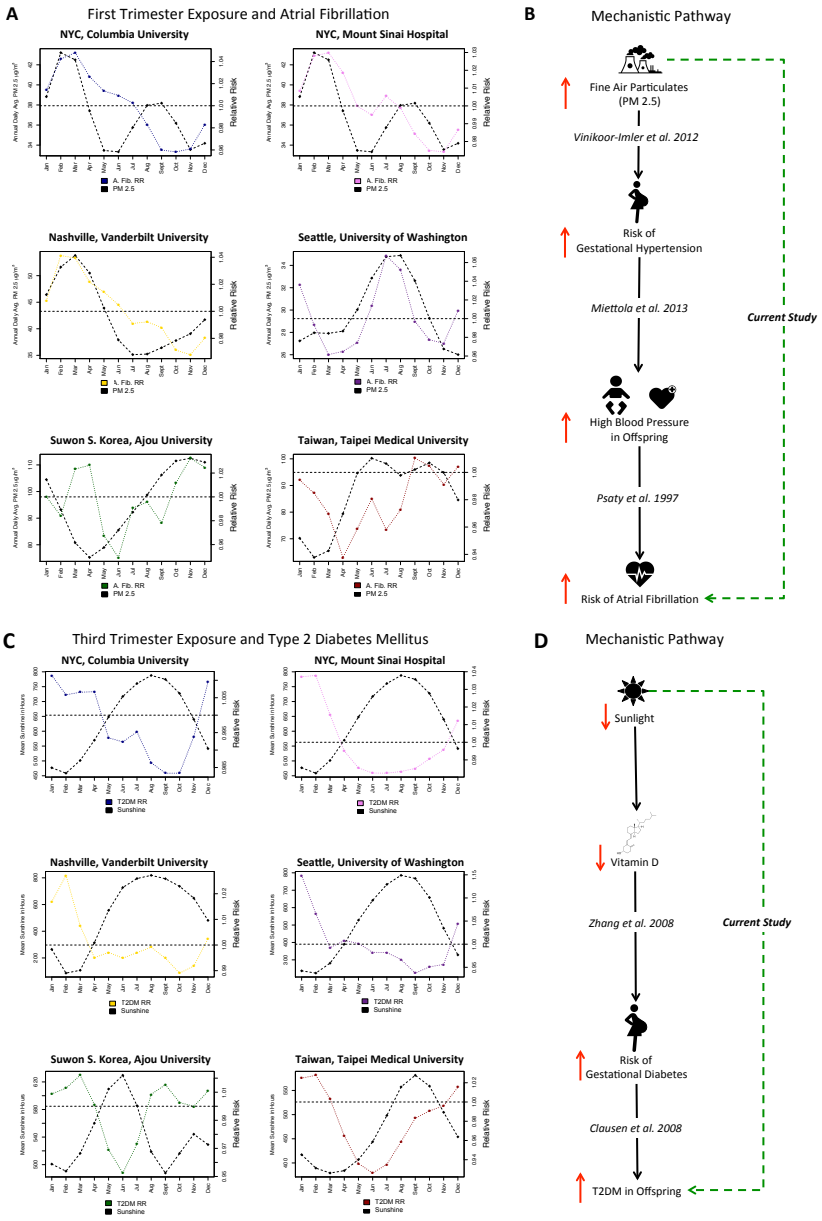


Figure 15. Atrial Fibrillation and First Trimester Exposure to Fine Particulate Matter (PM 2.5) and Type 2 Diabetes Mellitus and Third Trimester Exposure to Sunlight. Figure 15A. Atrial Fibrillation and First Trimester Exposure to Fine Particulate Matter (PM 2.5) At Each Study Site. The colored line is the relative risk of atrial fibrillation by birth month per site. Solid black lines indicate the first trimester exposure to PM 2.5 per site. **Figure 15B. First Trimester PM 2.5 Exposure and Offspring's Risk of Atrial Fibrillation: the Literature and Our Current Study.** A solid black arrow denotes each literature link with increase/decrease in risk depicted by up or down red arrows. I found a distal cause: prenatal exposure to PM 2.5 increases the risk of atrial fibrillation; whereas, others report findings of proximal causes in the same causal pathway. **Figure 15C. Type 2 Diabetes Mellitus (T2DM) and Third Trimester Exposure to Sunshine At Each Study Site.** The colored line is the relative risk of T2DM by birth month per site. Solid black lines indicate third trimester exposure to mean sunshine hours per site. **Figure 15D. Third Trimester Exposure to Sunshine and T2DM: the Literature and Our Current Study.** A solid black arrow denotes each literature link with increase/decrease in risk depicted by up or down red arrows. Low sunlight lowers vitamin D levels in the blood stream. My study is denoted by the green dashed arrow, which connected third trimester sunlight levels with T2DM risk later in life. Note that I uncovered the distal causal risk factors versus proximal causes.

A total of 56 distinct diseases were significantly associated with at least one exposure during at least one developmental stage. These 56 diseases were involved in 150 distinct disease-exposure-developmental stage tuples. Twenty-seven diseases were significantly associated across multiple exposure-stages. This was expected due to the inherent correlation among exposures. One disease – Dysuria – was involved in 14 tuples (disease-exposure-stage).

Several first trimester exposures were significantly correlated or anti-correlated with increased risk of depressive disorder later in life (**Figure 14A**), including low sunlight and temperature. However, the most significant association was a ***positive*** correlation between first trimester carbon monoxide (CO) exposure (R=0.725, Confidence Interval (CI): 0.529 – 0.847) and increased risk of depressive disorder. The relationship is shown in **Figure 14B** for all six individual sites**.**

Atrial fibrillation was positively correlated with PM 2.5 exposure during the first trimester (**Figure 15A**). Taiwan and South Korea both had fewer patients with atrial fibrillation (10,476 patients in Taiwan and 2,241 in S. Korea vs. US sites that ranged from 36,837 – 58,771 patients) and the relationship was not as strong in those locations (i.e., Taiwan and S. Korea). Further, lack of sunlight during both the third trimester and the perinatal period increased risk of T2DM later in life. The connection between low sunlight and increased risk of T2DM in the offspring was stronger during the third trimester (R=-0.816, CI: -0.5767 – -0.929) then during the perinatal period (R=-0.580, CI: -0.420 – -0.705). The individual site breakdown for the relationship between exposure to low amounts of sunlight during the third trimester and later risk of T2DM is shown in **Figure 15.**

3.5 Discussion

This study provides a global understanding of birth month - disease risk relationships and allows for systematic investigation of a number of different possible mechanisms. Results were integrated from over ten million unique individuals across three countries, two continents, and five distinct climates. This enabled successful discrimination between birth month – disease relationships driven by relative age (a cultural effect) versus seasonal environmental exposures, including climate factors, pollution and influenza. ADHD was found to be significantly dependent on relative age with an average difference in disease risk of 17.97% due to relative age with younger children experiencing greater risk of diagnosis than their peers. Several exposures were found to occur during the prenatal period (i.e., maternal exposures) that influenced risk of disease in the offspring and also perinatal exposures (i.e., direct exposure to the offspring) that influence lifetime disease risk.

3.5.1 Culture Effects Can Induce Birth Month – Disease Dependencies: The Tale of Relative Age

The 'relative age effect' is the phenomenon whereby children are preferentially selected based on their age relative to their peers (Helsen et al., 2005; Musch and Grondin, 2001). This is commonly studied among athletes where the slight advantage due to age including size, mental agility and timing of the onset of puberty provides slightly older children with a distinctive edge over their classmates. Sociologists have also looked into the effect and found that children that are 'younger' relative to their peers are at increased risk of being victims of bullying (Mühlenweg, 2010). Each of these relative age effects could alter an individual's risk of disease later in life. Therefore, my algorithm explicitly investigated the relationship between relative age, calculated using birth month distributions, and lifetime disease risk of all diseases in this

study. Of the 133 tested, only one disease – ADHD – was found to have a significant dependency due to relative age (**Figure 12**).

A study among Taiwanese children also found a significant relationship between relative age and ADHD (Chen et al., 2016). This effect was also found in Iceland, where in addition to finding a connection between relative age and ADHD, they also describe a relationship between academic performance and relative age (Zoëga et al., 2012). Other researchers have also studied the connection between academic performance, ADHD, and relative age finding increased risk for adverse outcomes among younger children (Evans et al., 2010).

While individual countries and sites have described the relationship between relative age and ADHD, this is the first comprehensive study to investigate relative age and disease across three distinct countries, six sites, and four distinct school cutoff dates. This increases the robustness of the findings while validating other site-specific studies, by establishing increased confidence in this dependency and the increased need for provider awareness of this issue when prescribing stimulants for the treatment of ADHD. This study establishes a relationship between ADHD and relative age, however the cause for this could be either a.) younger children are at increased risk of being a victim of bullying (which could result in traumatic brain injury, thereby affecting neurological functioning) or b.) over-diagnosis of younger children with ADHD (when they are otherwise healthy).

Note for the next three sections, all prenatal exposures affecting the offspring's lifetime disease risk are mediated through the fetal-maternal barrier. Therefore, all prenatal exposures are referred to as maternal exposures that influence the offspring's disease risk and all perinatal exposures are referred to as direct exposures on the newly born offspring.

3.5.2 Sunlight During Third Trimester and Risk of Type 2 Diabetes Mellitus in Offspring

Sunlight was inversely correlated with T2DM during the third trimester (R=-0.816) and the perinatal period (R=-0.580). Low vitamin D exposure during pregnancy has been linked to increased risk of gestational diabetes (Zhang et al., 2008), which is diabetes of the mother during pregnancy. Gestational diabetes is shown to increase the risk of T2DM among offspring with a reported odds ratio of 7.76 (Clausen et al., 2008). In this study, sunlight exposure during the third trimester of pregnancy was linked to changes in T2DM risk among offspring later in life.

The link between prior work described in the literature and results from my study is highlighted in **Figure 15D.** Each of the links details one small piece of the larger puzzle. In causal terms, these prior studies found the proximal causes whereas I reveal the distal factor that can explain the smaller steps in the proximal pathway (Laland et al., 2011; Scott-Phillips et al., 2011).

Mechanistically, my results also fit into the 'thrifty phenotype hypothesis', which states that inadequate early nutrition impairs development of the pancreas, which in turn greatly increases the susceptibility of the offspring to T2DM (Hales and Barker, 2013; Hales and Barker, 1992). Maternal gestational diabetes is often an indicator of impaired prenatal nutritional status. My work links a distal factor – low sunlight exposure during third trimester – to increased risk of T2DM in offspring. By uncovering a distal factor in this mechanism, this work opens the door for evolutionary biologists to delve into this relationship further to elucidate a fitness benefit.

3.5.3 First Trimester Exposures and Risk of Depressive Disorder in Offspring

Risk of depressive disorder and birth month is an association that is studied often in the literature (Joiner Jr et al., 2002). An Australian study investigated the relationship between birth month in both Southern and Northern hemispheres found that flu peak was important in explaining the birth season – depression/suicide relationship (Joiner Jr et al., 2002).

Figure 14A shows that first trimester exposure to influenza-like illness (ILI) was a significant

factor in depressive disorder with a slightly lower correlation value (R=0.612, CI: 0.384 – 0.770) than CO exposure (R=0.725, CI: 0.529 – 0.847). Depressive disorder was also significantly anti-correlated with sunlight (R=-0.625, CI: -0.452 – -0.753) and temperature (high temperature, R=-0.645, CI: -0.462 – -0.779; low temperature, R=-0.651, CI: -0.446 – -0.790) indicating that lack of sunlight during the first trimester also appeared to be related to depressive disorder. Because the strongest factor was CO exposure, I focused on the mechanisms underlying a relationship between first trimester CO and depressive disorder. Additionally, prior studies investigated a connection between ILI/flu and sunlight with depressive disorder without investigating pollutant variables such as CO that are often correlated with sunlight.

Generalized anxiety disorder (R=0.404, CI: 0.264 – 0.528) was also significantly associated with first trimester CO exposure although the relationship was weaker. Importantly, generalized anxiety disorder was only significantly associated with variance in CO exposure and no other variable (such as flu, sunlight, etc.). This further bolstered the hypothesis of a mechanistic link between both depressive disorder and generalized anxiety disorder and first trimester exposure to CO. Additionally, a study by Moffitt *et al.* found that generalized anxiety disorder (GAD) and major depressive disorder (MDD) often occur together with no apparent sequential pattern suggesting that GAD+MDD may be a disease of its own (Moffitt et al., 2007). Therefore, finding that both GAD and depressive disorder were significantly correlated with first trimester CO exposure suggests that this study may be uncovering a link between GAD+MDD and first trimester CO.

Chronic CO poisoning exhibits itself clinically as chronic fatigue, depression and often a diagnosis of influenza infection (either due to the patient's weakened immune system or due to the 'flu-like' symptoms that patients often present with) (Knobeloch and Jackson, 1999)

underscoring the importance of CO exposure on depression. Prenatal exposure to CO was shown to cause learning and memory deficits indicating that maternal exposure to CO crosses both the fetal-maternal and the blood-brain barriers (Mactutus and Fechter, 1984). First-trimester exposure to CO was shown to cause intrauterine growth retardation (IUGR) (Salam et al., 2005) and disrupts hippocampus functioning (Mereu et al., 2000). Shrinking of the hippocampus is one of the critical hallmarks of depression (Sapolsky, 2001). The link between first trimester CO and both GAD and depressive disorder may be mediated through a shrinking of the hippocampal structures caused by prenatal CO exposure. The link between this study and prior studies on prenatal CO exposure and depression from the literature is shown in **Figure 14C.** The major link in this study is the link between first trimester CO exposure and increased risk of depressive disorder (thick dashed green line). A lower correlation between GAD and first trimester exposure to CO was found suggesting that it could be patients afflicted with both diseases as described by Moffitt *et al.* (Moffitt et al., 2007) that were exposed to CO.

3.5.4 Fine Particulate Matter During First Trimester and Risk of Atrial Fibrillation in Offspring

A positive correlation was found between atrial fibrillation and PM 2.5 exposure during the first trimester (**Figure 15A**). Taiwan and South Korea both had very low incidence of atrial fibrillation and the relationship was not as strong in those locations suggesting the possibility that an additional factor may mediate the relationship. In adults, PM 2.5 exposure has been associated with adverse cardiovascular outcomes including increased heart failure admissions and mortality (Dominici et al., 2006; Ito et al., 2011; Zhou et al., 2011). Exposure to PM 2.5 in adults was also associated with increases in systolic blood pressure (Brook et al., 2010). Children of mothers with gestational hypertension were found to have higher blood pressure, and elevated cholesterol

and apolipoprotein B levels (Miettola et al., 2013). High blood pressure is a risk factor for later development of atrial fibrillation (Psaty et al., 1997). Exposure to fine air particulates increased the risk of gestational hypertension in pregnant women (Vinikoor-Imler et al., 2012). I propose a mechanism that connects atrial fibrillation and first trimester exposure to fine particulate matter by elevating maternal blood pressure and inducing gestational hypertension. The link between the prior literature on this topic and prenatal fine particulate matter and increased risk of atrial fibrillation is depicted in **Figure 15B.** The uncovered link between first trimester exposure to fine air particulates and increased risk of atrial fibrillation later in life represents a distal cause with the proximal causes all outlined together in **Figure 15B.**

3.5.5 Perinatal Exposures and Later Risk of Disease

Some diseases were tied to exposures during the perinatal period (i.e., the environment the baby is born into). One such relationship is perinatal flu exposure and lifetime risk of anemia (R=0.660, CI: 0.467 – 0.793). Some regions such as NYC and South Korea illustrated near perfect correlation between flu exposure and lifetime risk of anemia while other sites showed a lower correlation. Newborns are at increased risk from influenza and other viruses due to their developing immune system (Levy, 2007). This increases their risk of developing an infection due to the viruses. Anemia often results as part of the body's innate immune system to fight infections (Frickhofen et al., 1990). The uncovered link between perinatal flu exposure and anemia may be mediated through an immune pathway.

3.6 Limitations

My method investigates the presence or absence of correlations between exposures during different developmental stages and lifetime disease risk. The DSL meta-analysis method was used to uncover only correlations that were consistent across all study sites (i.e. robust). While

this study deeply probed into how specific exposures can affect lifetime disease risk, there are other exposures (e.g., diet) that this study was unable to investigate due to lack of available data. Importantly, if a co-varying environmental factor exists that was not included in this analysis and was ***perfectly correlated*** with one of the exposures in this analysis then my algorithm may be uncovering an association that is due to this other unmeasured factor. For example, if seasonal smoking patterns were perfectly correlated with CO seasonality at ***all six sites,*** then this would be a confounder. This is not likely given the number of sites (and perfect correlation of an environmental exposure across multiple sites is unlikely) and the diversity of the sites (e.g., Asian vs. USA). However, it remains a limitation of this study. Importantly, the phrase 'causal risk factors' is only used in instances where this study reveals the distal causal risk factor in an already established pathway. In other instances, further testing would be required to clearly state if this study's findings were causal factors or strongly correlated with another untested causal factor. Another limitation is unknown cultural differences between sites in my study. I investigated the relationship between relative age and birth month dependent diseases because the effect of relative age is possible to capture (i.e., through use of school cutoff dates). Other cultural differences may be difficult to identify and tease out (e.g., mandatory military conscription). Therefore, associations that are **not found across all sites** may be true birth month – diseases relationships where a cultural factor modulates the exposure. I focus only on those that are significant using the DSL method across all six sites.

3.7 Conclusion

In conclusion, this comprehensive study of factors involved in birth month – disease risk used data from over 10 million patients, three countries, two continents, and five climates. This study was able to distinguish the cultural effect of relative age from seasonal environmental exposures

that both affect birth month – disease dependencies. It also identified both the seasonal environmental exposure and the stage that resulted in increased disease risk. Others in the literature have identified the proximal causes behind these relationships, whereas this study identified distal causal risk factors. Several important findings include a link between both depressive disorder and generalized anxiety disorder and first trimester exposure to carbon monoxide. Lack of sunlight exposure during the third trimester was correlated with increased Type 2 diabetes mellitus risk. Finally, increased risk for atrial fibrillation occurred with first trimester exposure to fine air particulates. By identifying the distal causal risk factors in these disease pathways, this study allowed for the identification of areas that may require season-dependent dosing of prenatal supplements.

3.8 Acknowledgments

This chapter is a reproduction, in whole or in part, of a work submitted for publication (Boland et al., 2017c). I would like to thank Dr. Andrew Gelman, Department of Statistics, Columbia University for his tremendous help, support, and guidance during this project. I would also like to thank all co-authors on this paper.

Support for this research was provided through the following mechanisms: MRB was supported by the National Library of Medicine training grant T15 LM00707 (MRB) from Jul 2014 – Jun. 2016. MRB was supported by the NCATS, NIH, through TL1 TR000082, formerly the NCRR, TL1 RR024158 from Jul. 2016 – Jun. 2017. MRB and NPT were both supported by R01 GM107145. DS was supported by the National Library of Medicine training grant at the University of Washington T15 LM007442. SM was supported by the National Center for Advancing Translational Sciences (NCATS), NIH, through UL1 TR000423. SCY, RWP were supported by a grant of the Korea Health Technology R&D Project through the Korea Health

Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C0992).

Chapter 4

Measuring the Affect of Climate on Patient Mortality

4.1 Abstract

Climate is a known modulator of disease, but its impact on hospital performance metrics remains unstudied. The relationship between Köppen-Geiger climate classification and hospital performance metrics was assessed in this study. Specifically, 30-day mortality was obtained from Hospital Compare, and collected for the period July 2013 through June 2014 (7/1/2013 – 06/30/2014). A hospital-level multivariate linear regression analysis was performed while controlling for known socioeconomic factors to explore the relationship between all-cause mortality and climate. Hospital performance scores were obtained from 4,524 hospitals belonging to 15 distinct Köppen-Geiger climates and 2,373 unique counties. Model results revealed that hospital performance metrics for mortality showed significant climate dependence ($p<0.001$) after adjusting for socioeconomic factors. Climate is a significant factor in evaluating hospital 30-day mortality rates. These results demonstrate that climate classification is an important factor when comparing hospital performance across the United States.

4.2 Introduction

The relationship between climate and human health and disease is well known (Dell et al., 2012; 2013; Epstein, 1999; Hippocrates and Galen, 1952; Pöschl, 2005). Recently, variation in temperature related to climate variability and extreme weather events has been linked to hospital admission rates for heart disease (Schwartz et al., 2004). Air humidity is also an important health factor (Sherwood and Huber, 2010). Pollutants, including both carbon monoxide and fine air particulates have been associated with increased admissions for a number of conditions (Dominici et al., 2006; Schwartz, 1999). In contrast to the evidence (Lee et al., 2008), outcome-based quality of care assessments ignore hospital location when comparing hospitals and explicitly state the assumption that 'global location of the hospital should not affect the outcome' (Lacour-Gayet et al., 2004).

Hospital performance statistics are often reported while ignoring regional climate information. Location is usually only mentioned as a confounding factor in determining the fiscal cost of services (Hargraves and Brennan, 2016) while ignoring the quality of the result (i.e., the 30-day mortality rate). Many climate variables (e.g., temperature, humidity) and other climate-dependent pollutants (e.g., air particulates, water contaminants) may be responsible for some of the variation in hospital performance metrics. The total number of relevant climate variables to consider can be large. Therefore, the Köppen-Geiger climate classification system was selected (Köppen, 1884; Kottek, 2015) as a proxy for a high number of intertwined climate variables. The Köppen-Geiger climate classification uses precipitation and temperature to describe a region's climate and has been widely used in various fields such as hydrology(Peel et al., 2007) and ecology(Smith et al., 1992). In addition to climate, adjusting for socioeconomic factors is important when studying climate effects as prior studies have shown significant geographic

gradients in hospital services that are economic-dependent and not climate-dependent (Mercier et al., 2015).

Hospital performance statistics and 30-day mortality rates for many hospitals across the United States of America (USA) are readily available in the Hospital Compare dataset. Hospital Compare is maintained by the Centers for Medicare and Medicaid Services (CMS) with the purpose of providing the general public with information to make informed decisions on their healthcare (www.hospitalcompare.hhs.gov).

In this study, hospital-level variation in 30-day mortality measures reported in Hospital Compare and their relationship with hospitals' Köppen-Geiger climate classification were examined.

4.3 Methods

4.3.1 Data

4.3.1.1 Köppen-Geiger climate classification

This study used the Köppen-Geiger climate classification system (Köppen, 1884; Kottek et al., 2006) to compare hospitals' climates. The Köppen-Geiger classification is a classical climate classification used by researchers (Zanobetti and Schwartz, 2009) throughout the world including members of the World Health Organization (Polack et al., 2005). Each county in the United States of America (USA) is assigned a category using three axes: broad climate type, as well as precipitation and temperature characteristics (Köppen, 1884; Kottek, 2015). Each of these three factors is combined to produce the overall climate designation. For example, the New York City (NYC) climate is designated Cfa meaning that its climate is warm temperate ('C'), fully humid ('f') with a hot summer ('a'). Köppen-Geiger climates at the US-county level were obtained (Kottek, 2015) and linked it to Federal Information Processing Standard (FIPS) codes. In some

instances, multiple climate classes were mapped to the same county with proportions for each climate within that county (this occurs when counties contain data from across two climate boundaries). Therefore, the climate with the highest proportion for each county (i.e., the dominant climate) was used in this study's analysis.

4.3.1.2 Hospital Compare

Thirty-day mortality data from Hospital Compare were obtained by downloading the entire release of Hospital Compare (2015 Annual Files released on July 16, 2015) (CMS, 2015). The 'readmissions, complications and deaths' file was used to obtain 30-day mortality rates for all conditions reported (a total of 6 conditions, obtained from file: HOSArchive_Revised_FlatFiles_20150716/Readmissions and Deaths-Hospital.csv). Data for these measures were collected for the period from July 2011 through June 2014 (7/1/2011 – 6/30/2014). This study did not investigate patient-reported outcomes (Boulding et al., 2011). Each hospital included in Hospital Compare contains zip code information, which can be used to link the hospital to the county-level Köppen-Geiger climate data source (described above).

4.3.1.3 Census Bureau's American Community Survey (ACS)

Socioeconomic factors are known confounders in quality-of-care comparisons across hospitals (Das and Mohpal, 2016). Therefore, data on six different potentially confounding variables for hospital performance were obtained, including: income, race, English-speaking ability, insurance coverage, renter-occupied status, and total number of households. The American Community Survey (ACS) collected by the U.S. Census Bureau, 5-year data from the 2014 release was used.

These six potential confounders were selected because they have been linked to hospital performance (either readmission or mortality) metrics and are known to vary regionally. Socioeconomic status (specifically median household income and race) are well-known

independent risk factors for hospital readmissions related to heart failure (Philbin et al., 2001). Patients who owned their own homes had significantly fewer hospital readmissions for COPD (Coventry et al., 2011) indicating that the percent of renters per county is a potential confounder. English fluency is another potential confounder as non-English speaking individuals were at an increased risk for 30-day readmissions even after adjusting for other socioeconomic variables (Karliner et al., 2010). Insurance status also altered risk for hospital readmissions (Kangovi and Grande, 2011).

4.3.2 Statistical Methods

I extracted all six mortality metrics for each hospital provided by Hospital Compare. Each score was reported in percentages. For example, a hospital score of 14.6 for ’Heart Failure (HF) 30-Day Mortality Rate” indicates that 14.6% of patients with a heart failure diagnosis died within 30-days. Raw mortality counts were used along with the counts of patients sampled for each hospital instead of the raw mortality rates. I mapped the Hospital Compare data to their corresponding FIPS codes using the zip codes provided in the original data file using the R package noncensus (Ramey, 2015).

The Köppen-Geiger climate classes were mapped to hospital county information contained in the Hospital Compare data. All scores that were reported as ‘Not Available’ by the CMS were removed. The CMS lists data as ‘Not Available’ if 1) the number of cases does not meet the minimum required for public reporting, 2) the number of cases is too small to reliability report the data, or 3) to protect personal health information. The CMS may also restrict access to mortality rates if 1) a hospital elected not to submit data for a particular reporting period, 2) a hospital had no claims data for a particular measure or 3) a hospital elected to suppress a measure from being publicly reported. Since the CMS chose to restrict these data, they were

unavailable for analyses. I also removed four hospitals belonging to a unique climate (i.e., with no other hospital in that climate) to avoid any biases due to low-sample size. The final dataset contained 4,524 hospitals belonging to 15 distinct climates. The F-test was used to determine overall significance for the various measures and confounders vs. climate.

A linear hierarchical regression model was constructed that pools data from all six 30-day mortality rates reported in Hospital Compare. This model was implemented using the Stan software in R (Carpenter et al., 2016). The model then automatically captures the relationship between size of the hospital and natural sampling fluctuations. The model was fitted using the open-source Bayesian inference engine, running at the default settings of 2000 iterations for 4 chains. Parameters for hospital size, and hospital-specific variability in mortality were used in the model as these factors could skew the results (Gelman and Price, 1999). In addition, parameters for each of the six-socioeconomic confounder variables were used. These socioeconomic factors included: income, total number of households, % renter, % un-insured, % speak English 'very well', and % white alone. Data for each socioeconomic variable were obtained from the ACS using the county-level FIPS codes. Therefore, hospitals were assigned their corresponding socioeconomic variables based on the geographic location of the hospital (i.e., all hospitals in the same county received the same set of socioeconomic variables).

4.4 Results

The final dataset contained 4,524 hospitals from 15 distinct climates and 2,373 unique counties (**Figure 16**). All six mortality measures varied by climate as illustrated in **Figure 17** and the six socioeconomic confounder variables also varied across climates and by USA county (**Figure 18**). Each confounder was significant across different climates, which motivated the addition of adjustment for these socioeconomic confounders in the model.

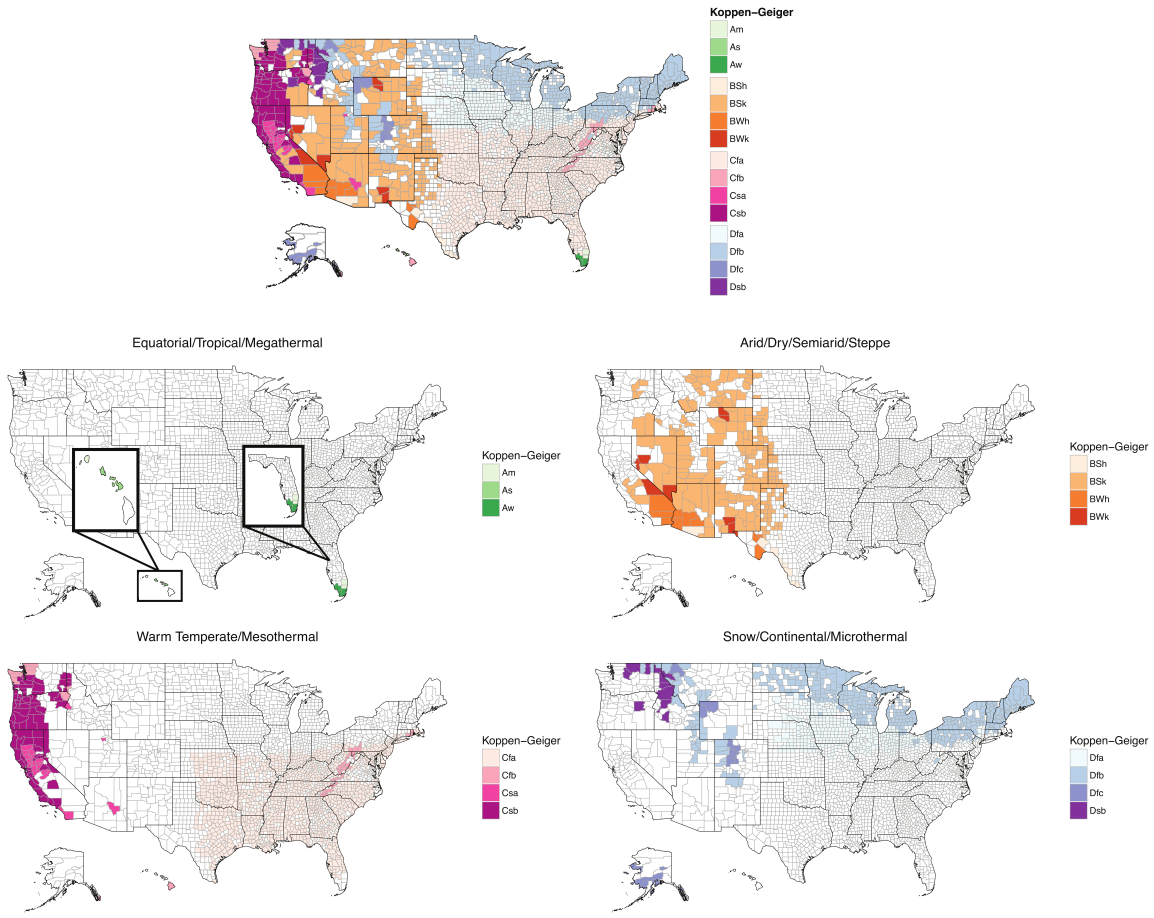


Figure 16. Hospital Compare Data By County with Major Climate Designations: A Map of the United States Showing Hospital Compare Data Mapped to Köppen-Geiger Climate Classifications. Map of the United States was generated using the following R libraries: choroplethr (version: '3.5.2', URL: https://cran.r-project.org/web/packages/choroplethr/index.html), ggplot2 (version: '2.1.0', URL: https://cran.r-project.org/web/packages/ggplot2/index.html), noncensus (version: '0.1', URL: https://cran.r-project.org/web/packages/noncensus/index.html), zipcode (version: '1.0', URL: https://cran.r-project.org/web/packages/zipcode/index.html), grid (version: '3.3.0', URL: https://stat.ethz.ch/R-manual/R-devel/library/grid/html/00Index.html) and gridExtra (version: '2.2.1', URL: https://cran.r-project.org/web/packages/gridExtra/index.html). The map itself utilized the choroplethr library version 3.5.2.

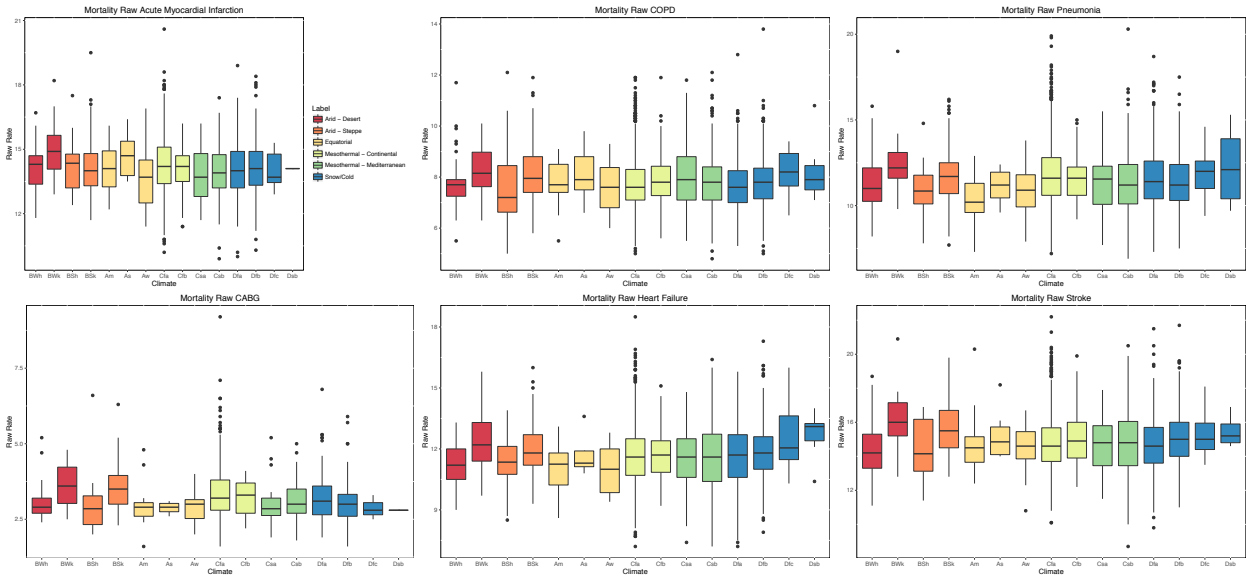


Figure 17. Raw Mortality Boxplots For All Six Mortality Measures By Köppen-Geiger Climate Classification System. The relationship between climate and individual disease varies somewhat. A near linear relationship is observed for 30-day heart failure mortality (lower center plot). A pooled-mortality statistic (across all 6 diseases) was used in the model. Köppen-Geiger Model data obtained from (Köppen, 1884; Kottek, 2015).

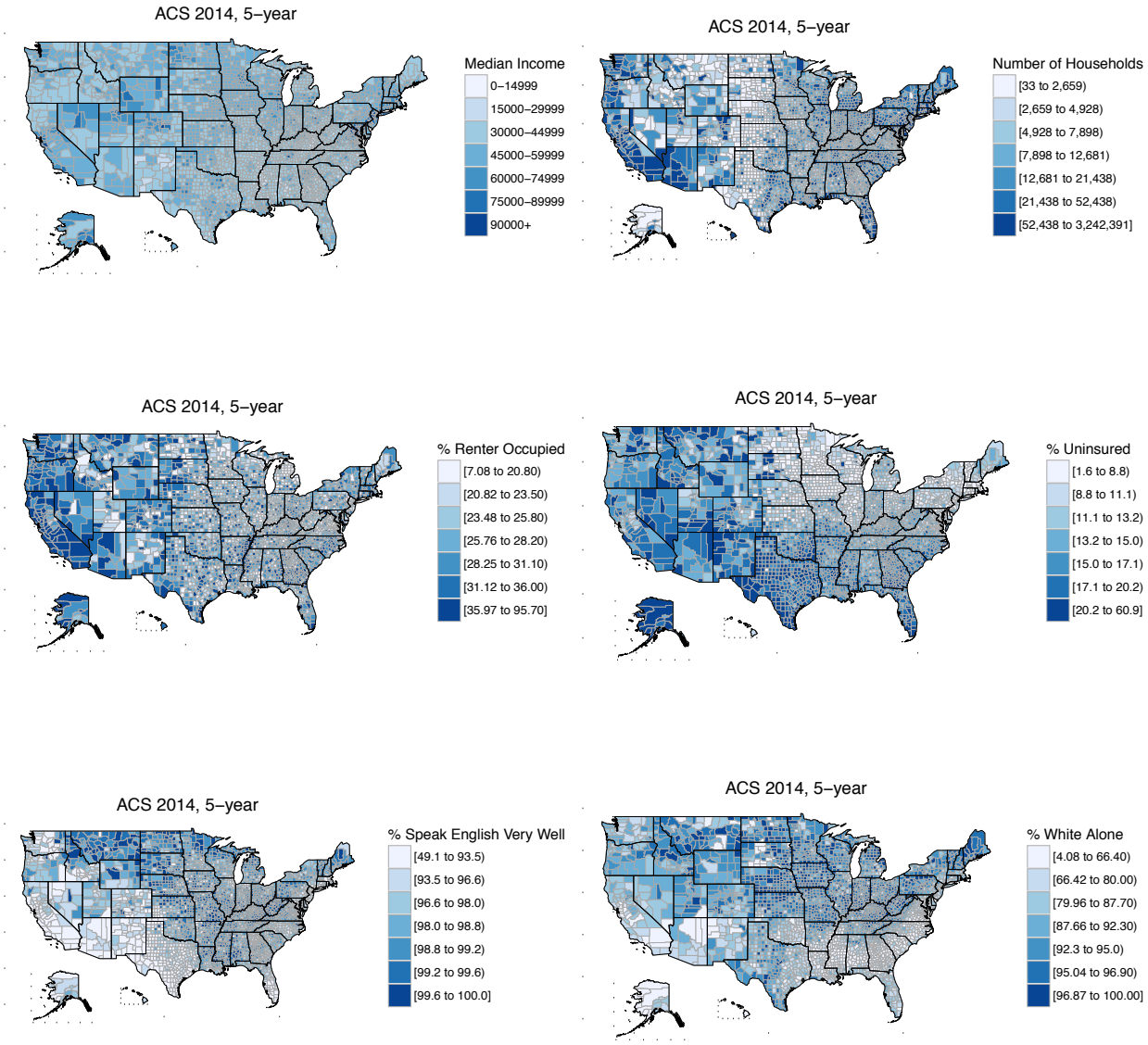


Figure 18. County-Level Variance of Six Known Confounders: income, total number of households, % renter occupied housing, % uninsured persons, % English-fluent and % white. Map of the United States was generated using the following R libraries: choroplethr (version: '3.5.2', URL: https://cran.r-project.org/web/packages/choroplethr/index.html), ggplot2 (version: '2.1.0', URL: https://cran.r-project.org/web/packages/ggplot2/index.html), noncensus (version: '0.1', URL: https://cran.r-project.org/web/packages/noncensus/index.html), zipcode (version: '1.0', URL: https://cran.r-project.org/web/packages/zipcode/index.html), grid (version: '3.3.0', URL: https://stat.ethz.ch/R-manual/R-devel/library/grid/html/00Index.html) and gridExtra (version: '2.2.1', URL: https://cran.r-project.org/web/packages/gridExtra/index.html). The map itself utilized the choroplethr library version 3.5.2.

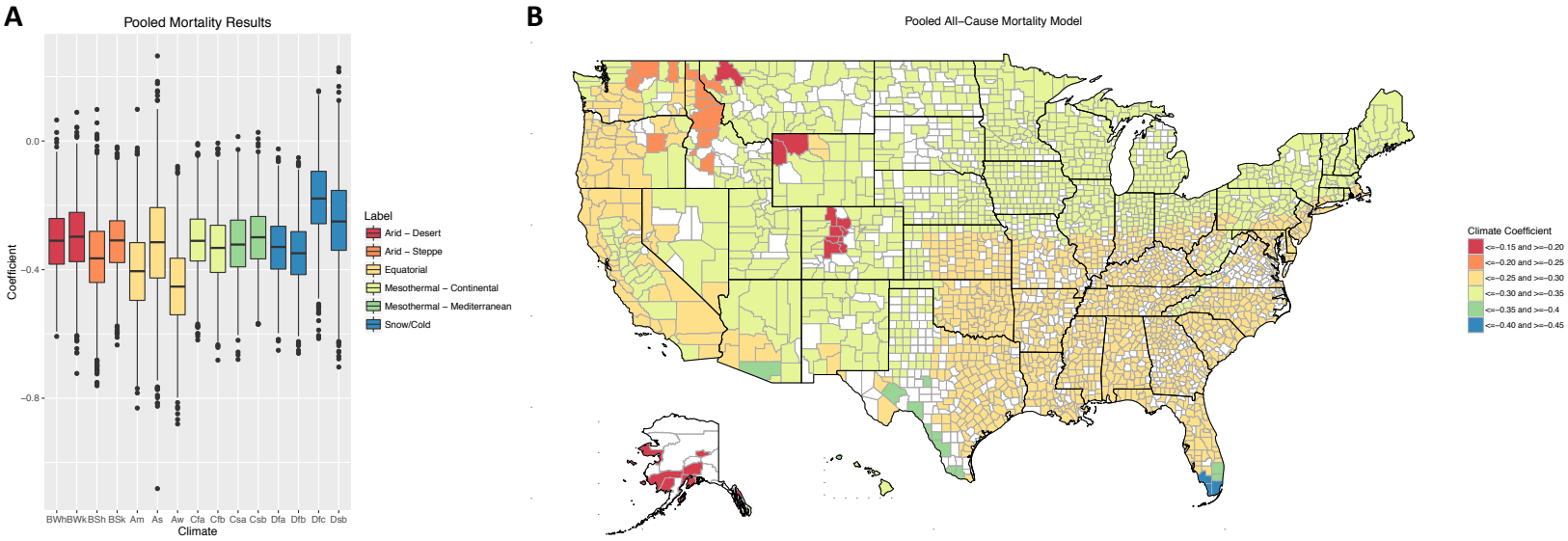


Figure 19. Climate's Impact on Hospital Performance Mortality Statistics After Adjustment for Confounders: Map of the United States of America. Model coefficients for climate's impact on 30-day mortality are displayed by climate classification in **Figure 19A**. A map of the USA illustrating the results of the model is shown in **Figure 19B.** Map of the United States was generated using the following R libraries: choroplethr (version: '3.5.2', URL: https://cran.r-project.org/web/packages/choroplethr/index.html), and ggplot2 (version: '2.1.0', URL: https://cran.r-project.org/web/packages/ggplot2/index.html). The map itself utilized the choroplethr library version 3.5.2.

The linear regression model revealed distinct relationships between climate and mortality statistics. The mortality data was pooled across all six conditions and found that equatorial climates (light orange **Figure 19A**) had lower climate coefficients after adjustment for confounding. Hospitals located in equatorial dry climates (Aw climate) were found with the lowest risk of mortality (blue in **Figure 19B**) indicating that wetter climates are more debilitating a finding also reported by Sherwood *et al.*'s model (Sherwood and Huber, 2010). The subarctic snow climate (Dfc climate) had the highest pooled mortality rate (red in **Figure 19B**). Another climate with high mortality is the snowy warm summer climate (orange in **Figure 19B,** Dsb climate). Both the subarctic snow climate (Dfc) and the snowy warm summer climate (Dsb) are centered in the northwest region of the USA.

4.5 Discussion

This study's results suggest the importance of considering climate-induced impact in hospital performance statistics. Pooled 30-day mortality rates were found to vary significantly by climate. In general, equatorial climates had improved mortality rates after adjusting for socioeconomic confounders. Further, subarctic and heavy snow climates had worse mortality rates. The model adjusted for known confounders, including, % insured, % renters, household income, % speak English well, % white and total number of households. The model found a significant climate effect after adjusting for these factors. The model's results fit well with the literature. For example, cold, dry air (i.e., low temperature and low humidity) is known to increase the risk of influenza-related mortality (Barreca and Shimshack, 2012; Davis et al., 2012) and my model found that colder climates tended to have higher mortality rates while warmer, milder climates had lower mortality rates. Additionally, Sherwood *et al.* demonstrate that peak heat stress for humans and animals occurs in wet climates indicating that dry climates may be the ideal

environment while high temperature and relative humidity can be debilitating (Sherwood and Huber, 2010).

4.5.1 Policy Implications Of Climate – Performance Relationship

Government agencies, such as CMS, are making tremendous strides to provide more data to patients to allow patients to make informed healthcare choices. The CMS designed Hospital Compare as a patient-facing tool to allow patients to 'compare' hospitals using a set of hospital performance metrics. This allows patients to decide where they want to receive medical treatment.

Some researchers have questioned the usefulness of Hospital Compare's performance metrics (Halasyamani and Davis, 2007). They note a discrepancy between "Hospital Compare" quality measures and "Best Hospitals" (Halasyamani and Davis, 2007). However, researchers have shown that the reported conditions (heart attack, pneumonia, heart failure) that Hospital Compare highlights account for 16% of Medicare discharges from acute care hospitals and 16% of Medicare hospital payments (Kahn et al., 2006). This makes the Hospital Compare performance metrics vital for a large cohort of Medicare patients.

Additionally, alternative metrics such as "Best Hospitals" have not proven to be informative in predicting outcomes such as 30-day mortality. One study investigated surgical outcomes following radical cystectomy using "Best Hospitals" performance metrics from the US News & World Health Report (Lascano et al., 2015). The researchers found either no correlation or an inverse correlation between the quality of the hospital (using one of these "Best Hospitals" metrics) and mortality 90 days following surgery (Lascano et al., 2015) making those metrics not informative for measuring mortality-related hospital performance. Others found that admission to one of "America's Best Hospitals" was associated with lower 30-day mortality among elderly

patients with acute myocardial infarction (Chen et al., 1999). Although, those researchers did not mention adjusting for other socioeconomic confounders (e.g., insurance status) that would bias their results (Chen et al., 1999).

Hospital Compare performance measures have been used successfully by many researchers to learn more about hospital care through the United States (Werner and Bradlow, 2006). Using these data, this study demonstrates that climate matters when choosing a hospital even after adjusting for many other known socio-economic factors (**Figure 19B**). These results have important implications for patients as well as policy makers.

4.5.2 Proposed Climate-Based Performance Adjustment

Originally, hospitals' mortality rates were compared against a national rate. However, due to unfair penalization of hospitals in poor-regions, adjustments were made for socioeconomic status (Glance et al., 2015). In this study, climate was found as another factor in hospital mortality rates and therefore I propose the notion that financial adjustments should be made for hospitals within their own climate. This would allow hospitals to be meaningfully compared to each other. Climate-based adjustments should be considered by policy makers to make meaningful comparisons across different hospitals.

4.6 Limitations

There are some limitations to this study including the exclusive use of Hospital Compare's 30-day mortality metrics. These metrics do not fully represent the quality of a given hospital (Zuger, 2015). However, these metrics are used to assess hospital quality by policy makers and these mortality measures represent a significant burden on the Medicare population (Kahn et al., 2006). 30-day mortality rates were used because they were readily available from Hospital

Compare. However, some researchers suggest that shorter time intervals may be more conducive to quality-of-care assessments (Chin et al., 2016), however these shorter time intervals are not currently being provided publically.

4.7 Conclusion

This study demonstrates that climate-induced impact on hospital performance metrics exists. This climate-based variation in mortality rates exists among hospitals even after adjusting for socioeconomic confounders. The findings are important for policy makers as climate-based adjustments for hospitals could be conducted to enable 'meaningful comparisons' of hospitals by comparing hospitals within the same climate. Additionally, this study's findings are important for researchers that study hospital performance as the Köppen-Geiger climate class where a hospital is located represents an important, yet often overlooked, variable in the equation.

4.8 Acknowledgments

This chapter is a reproduction, in whole or in part, of a work submitted for publication (Boland et al., 2017b). I would like to thank Dr. Andrew Gelman, Department of Statistics, Columbia University for his tremendous help and support when designing the statistical model and with assistance in implementing the algorithm using the STAN software suite. I would also like to thank all co-authors on this paper. Support for this research provided by R01 GM107145 (MRB, NPT). MRB was supported by the National Library of Medicine training grant T15 LM00707 (MRB) from Jul 2014 – Jun. 2016. MRB was supported by the NCATS, NIH, through TL1 TR000082, formerly the NCRR, TL1 RR024158 from Jul. 2016 – Jun. 2017.

Section II

Mechanistic Insights

Chapter 5

Uncovering Genes Underlying Birth Season – Disease Effects

5.1 Abstract

Prenatal and perinatal exposures vary seasonally (e.g., sunlight, allergens) and many diseases are linked with variance in exposure. Epidemiologists often measure these changes using birth month as a proxy for seasonal variance. Likewise, Genome-Wide Association Studies have associated or implicated these same diseases with many genes. Both disparate data types (epidemiological and genetic) can provide key insights into the underlying disease biology. In this chapter, I developed an algorithm that links 1) epidemiological data from birth month studies with 2) genetic data from published gene-disease association studies. My framework uses existing data repositories – PubMed, DisGeNET and Gene Ontology – to produce a bipartite graph that connects enriched seasonally varying biofactors with birth month dependent diseases (BMDDs) through their overlapping developmental gene sets. As a proof-of-concept, I investigate 7 known BMDDs and highlight three important biological graphs revealed by my

algorithm and explore some interesting genetic mechanisms potentially responsible for the seasonal contribution to BMDDs.

5.2 Introduction

Dopico *et al.* demonstrated that gene expression can vary seasonally (Dopico et al., 2015). Additionally, many biological compounds are known to vary seasonally in humans (Basu et al., 1994; Grzybowska et al., 1993; Hao et al., 2003; Woodhouse and Khaw, 2000). To address aim three, an algorithm was needed that uses existing data repositories containing genetic information for various disease states and Seasonally Varying Biofactors (SVBs) to find genes potentially responsible for reported birth month dependent diseases (BMDDs). These genes should lend insight into the underlying mechanisms behind birth month-disease relationships.

One well-studied BMDD is asthma. Several studies have linked asthma risk to birth month (Boland et al., 2015b; Korsgaard and Dahl, 1983) where birth month is a proxy for a perinatal environmental exposure. Environmental factors play a key role not only in asthma development but also in its progression. Others have demonstrated that asthma flare-ups are seasonally dependent (Cohen et al., 2014; Randolph, 2014) and that genetic mechanisms are involved in this seasonal dependency (Bjornsdottir et al., 2011). Despite all the knowledge behind asthma seasonality and the role of perinatal exposures with regards to disease risk, it remains impossible to identify a biological mechanism behind the birth month association. In part, the difficulty lies in the fact that >1,200 genes have been implicated in asthma disease progression (Piñero et al., 2015). Because of the plethora of genes implicated in certain diseases, finding the potential genetic mechanisms underlying birth seasonality associations is non-trivial. This formed the motivation for aim three.

Aim three focuses on constructing a Month-Of-Birth (MOB) Gene map (MOB-Gene map) that

uses publically available data sources including PubMed and DisGeNET. The method builds on the traditional Mendelian Randomization (MR) Approach (**Figure 20**) with several important modifications.

The traditional MR approach (Smith and Ebrahim, 2003) involves studying the relationship between an exposure (e.g., lead paint) and a disease (e.g., asthma) through the use of an intermediary proxy – namely, genetic polymorphisms. The idea is that if certain gene polymorphism changes occur following exposure to a certain substance and those same gene polymorphisms changes also occur in the presence of the disease then there may be a link between the two entities via the gene polymorphism changes. This can also be statistically quantified using enrichment analysis. However, there are several limitations with this approach because a disease resulting from an exposure (e.g., colon cancer) can have very different properties than the type of colon cancer that is inherited via gene polymorphism changes.

The MOB-Gene map required a link between a seasonally varying biofactor (e.g., vitamin D) and various diseases or BMDDs. The process was similar but instead of gene polymorphism changes, any gene link was used (i.e., gene must be connected to the disease and SVB in DisGeNET). Additional restrictions were placed on the gene to only include genes that were expressed during development. This results in a MOB-Gene Map where BMDDs are linked with SVBs using their underlying genes (Boland and Tatonetti, 2016c).

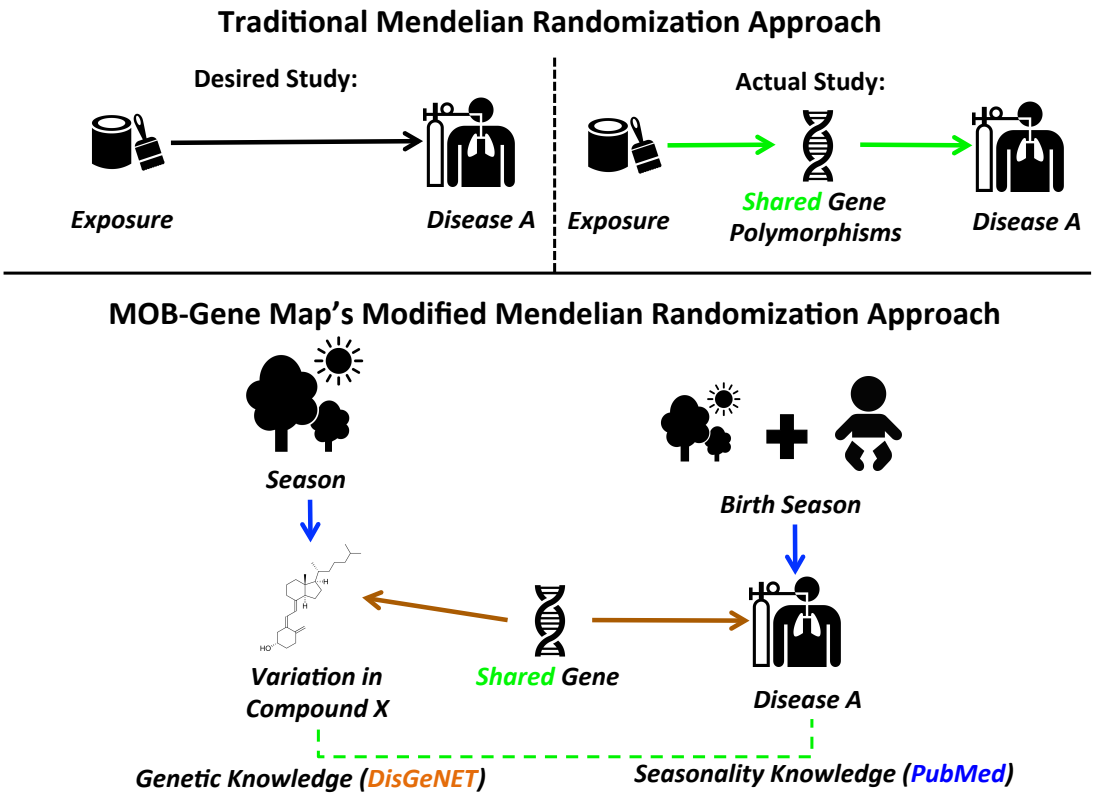


Figure 20. Differences Between The Traditional Mendelian Randomization Approach And My Approach. The Traditional Mendelian Randomization Approach Uses Shared Gene Polymorphism Data. My Modified Mendelian Randomization Approach Uses Shared Genes Involved in Seasonally Varying Biofactors and Disease.

5.3 Methods

My framework combines data from three public data repositories: PubMed (http://www.ncbi.nlm.nih.gov/pubmed), DisGeNET (http://www.disgenet.org) (Piñero et al., 2015) and the Gene Ontology (GO - http://geneontology.org). **Figure 21** illustrates the overall framework approach. Each step is described in more detail in the sections that follow.

5.3.1 Assembling Data Sources from Existing Data Repositories

Using PubMed, my algorithm searched for SVBs in humans (*Homo sapiens*). All non-humans (e.g., rats, geese, and even non-human primates) were excluded and human physiological state was ignored (e.g., post/pre menopausal, old/young). **Figure 22** contains an example of two SVBs extracted from a study by Meier et al. (Meier et al., 2004) demonstrating the seasonality of parathyroid hormone (PTH) and vitamin D (specifically calcifediol). In **Figure 22**, one can see that PTH tends to be higher in the winter months (Jan-Mar) while vitamin D is noticeably higher in the late spring / summer months (Jun-Aug).

To develop a list of literature-backed SVBs, I first queried PubMed using the following:

```
(human) AND "seasonal variation"
```

which returned 4,091 articles. I then added a species filter (humans) and a language filter (English), which reduced the results set down to 3,627 articles.

This study is primarily interested in SVBs and not disease flare-ups (e.g., asthma exacerbations occur seasonally). Therefore, I modified the query to also include the compound. I then ran this for a large variety of compounds (e.g., vitamin D, lactic acid, eosinophils, neutrophils, estrogen, testosterone) to retrieve articles related to their seasonality or lack thereof. I read through the resulting abstracts to determine if the result was correct and to remove any non-human studies

that passed through the earlier filter. After these initial checks, I determined if seasonal variation was found or not found by the study.

Previously, a curated reference set of BMDDs was assembled to assess the quality of SeaWAS results (Boland et al., 2015b). The details of the assemblage of this set is described in chapter two and in the subsequent literature publication (Boland et al., 2015b). Because I wanted a list of BMDDs with at least 1 publication supporting the relationship between birth month and disease risk, the original list was extended to include the 12 novel findings from my SeaWAS study described in chapter two. Supplemental information is available for this study on figshare. I will use the phrase 'see supplement' throughout this chapter when I am referring to figshare accessible via the following link:

figshare.com/s/b47610ea62d111e5b56406ec4bbcf141

Each SVB was mapped to a disease involving dysregulation of a SVB because DisGeNET only contains genes associated with disease states. For example, vitamin D is an SVB. Hence vitamin D deficiency was used as one of the diseases involving the SVB vitamin D. All genes implicated via association studies in the disease of vitamin D deficiency were used as vitamin D genes. To match SVBs to diseases in DisGeNET, substring matching was performed using the SVB query term. In some cases, I had to modify the SVB search term used. This was done to ensure that SVBs such as vitamin C (also known as ascorbic acid) were mapped properly. A list of SVBs and the exact query terms used for extracting genes from DisGeNET is included with the supplement. Examples in **Table 9** are included for explanatory purposes along with example diseases and example genes implicated in those diseases for each SVB in **Table 9.** However, these examples are not exhaustive.

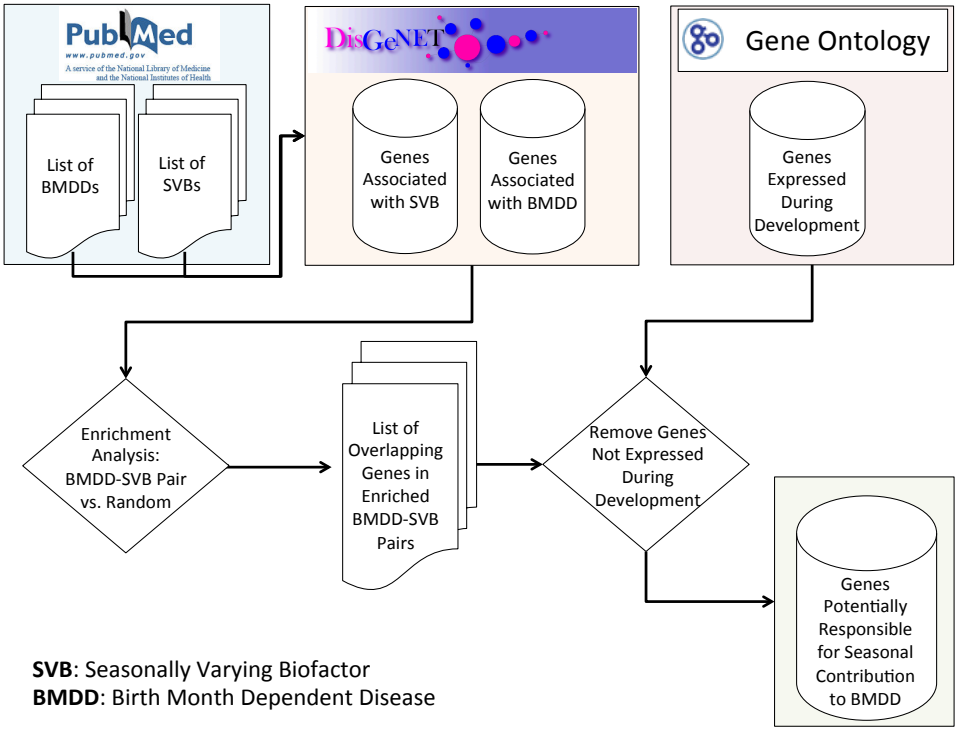


Figure 21. Overview of My Method Designed to Locate Genes Potentially Responsible for Seasonal Contribution to BMDD.

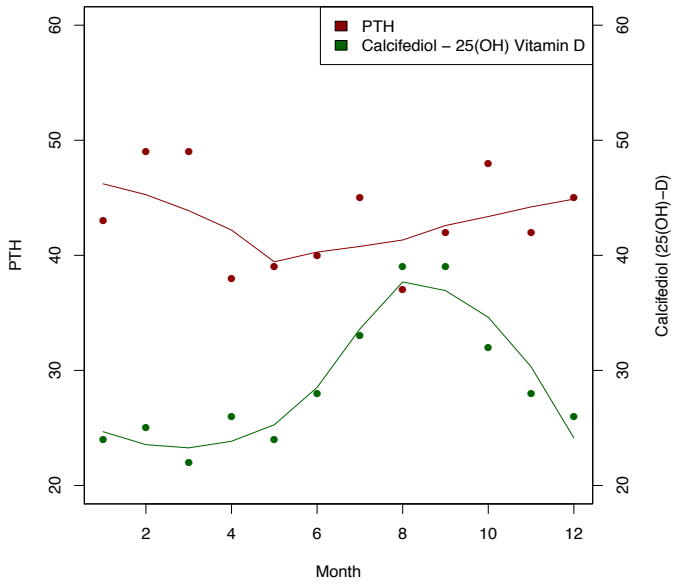


Figure 22. SVBs (parathyroid hormone and calcifediol) measured by Meier et al. 2004. Best-fit lines were applied and slight non-significant anti-correlation was observed (r=-0.303, p=0.338).

Table 9. Examples of SVBs and DisGeNET query terms used to extract SVB-related diseases and genes potentially involved in perturbation of SVBs

SVB	SVB query term used	Example Disease Names	Example Genes Implicated
Vitamin D	"Vitamin_D"	Vitamin D Deficiency, Rickets Hereditary Vitamin D-Resistant	DHCR7, VDR
Parathyroid hormone	"Parathyroid"	Pseudo-hypoparathyroidism, Parathyroid Neoplasms	GNAS, CDC73
Vitamin C	"ascorbic_acid"	Ascorbic Acid Deficiency	GSTK1, HP, SLCO6A1
Vitamin K	"Vitamin_K"	Vitamin K Deficiency, Vitamin K Dependent Clotting Factors Combined Deficiency	GGCX, VKORC1, F7
Neutrophil	"Neutrophil"	Neutrophil Actin Dysfunction, Hereditary Neutrophilia	PARP1, CYBA, CSF3R
Eosinophil	"Eosinophil"	Eosinophilia, Hyper-eosinophilic Syndrome	IL5, FIP1L1, PDGFRA
Hemoglobin	"Hemoglobin"	Hemoglobinopathies, Methemoglobinemia	HBB, CYP1A2, HBA1
Estrogen	"Estrogen"	Oestrogen deficiency, Estrogen Resistance	ESR1, RBBP4, BCAR1, PPP2CA

Additionally, I mapped the list of BMDDs to diseases in DisGeNET. I did this using approximate string matching similar to above. For this proof-of-concept, I randomly chose 7 BMDDs for this analysis provided they spanned the distribution of the number of distinct genes. The seven BMDDs and their query terms used in DisGeNET are given in **Table 10**. I only used 7 BMDDs because I wanted to test the feasibility of my framework and algorithm. Note the number of distinct genes involved in each disease varies largely from 21 genes (reproductive performance) to 1253 genes (asthma). I randomly selected the BMDDs with this one constraint.

Table 10. BMDDs included in proof-of-concept along with query terms, example genes implicated and counts of genes involved in BMDD

BMDD	BMDD query term used	Example Genes Implicated	No. Distinct Genes
Asthma	"Asthma"	SCGB1A1, TNF, CCL11	1253
Attention Deficit Disorder	"attention_deficit_disorder"	COMT, LPHN3, GRM5	338
Atrial Fibrillation	"Fibrillation"	ACE, NOS3, KCNE2, SELE, VWF	318
Reproductive Performance	"Reproductive"	BRCA1, BRCA2, TLR4, ESR1, MBL3P	21
Cardiovascular Disease	"Cardiovascular"	ACE, APOB, LPL, MTHFR	775
Cardiomyopathy	"Cardiomyopathy"	CSRP3, TTN, DES, TMPO, VCL	717
Mitral Valve Disorder	"mitral_valve"	FBN1, AGTR1, FBN2, NPPB, PLAU, COL3A1	82

5.3.2 BMDD-SVB Pair Enrichment Algorithm

My algorithm is designed to uncover BMDD-SVB pairs that were enriched using their respective gene sets. The first step was the creation of an empirical null distribution for each disease. For each BMDD, genes were randomly extracted from DisGeNET (of the same size as the number of

genes for that BMDD). Therefore, 1,253 distinct genes would be randomly pulled from DisGeNET for the empirical null distribution for asthma. However, for reproductive performance only 21 genes would be randomly pulled. The overlap was calculated between the SVB and the random gene set. This sampling protocol was iterated for 100 times and then an overall average overlap was computed. Fisher's exact test was performed between the true BMDD-SVB pair and the random average overlap computed above (**Table 11**). Bonferroni correction was applied to adjust for multiple comparisons.

5.3.3 Restrict to Genes Involved in Developmental Processes

Using the Gene Ontology (GO), the gene set was restricted to only include those genes involved in developmental processes. Only genes with at least one GO term containing 'develop' in its annotation term description were retained. So for example, if a gene contained the term 'positive regulation of hair follicle development' or 'embryonic placenta development' it was retained. This further reduced the gene set to about 30% of the size (for asthma, 439 asthma genes were enriched in asthma-SVB pairs and only 140 had at least one developmental GO term).

5.3.4 Construct Bi-Partite Networks

To visualize the results, I created bi-partite graphs similar to (Lorberbaum et al., 2015) for each SVB. BMDDs are included if they are enriched for overlapping genes with the SVB of interest. Each SVB (shown on the left side of the graph) is linked to BMDDs (shown on the right side of the graph) that are enriched in overlapping genes that are depicted in the middle portion of the graph. Only genes with a developmental GO process are included. DAVID was used (Huang et al., 2008; 2009) to annotate the genes and identify functional gene modules. Network visualization was performed using Cytoscape (Shannon et al., 2003).

Table 11. The Structure of the Enrichment Algorithm: Each BMDD-SVB Pair was Compared Against a Randomly Generated BMDD-SVB Pair Specific for that BMDD

	No. of BMDD Genes Per SVB	No. Genes Per SVB – No. BMDD Genes Per SVB
Actual BMDD-SVB Pair	**A**	**B**
Randomly Generated*	**C**	**D**

*** Random was the average across 100 random gene set extractions from DisGeNET using the same number of genes as the BMDD**

5.4 Results

5.4.1 Using Existing Data Repositories to Assemble Key Datasets

The original search returned 3,627 articles related to seasonal variation in humans. Therefore, I included additional query terms for biological compounds such as hormones, vitamins, and immune-related cells that are thought to vary seasonally. This allowed for the identification of 22 SVBs that are known to vary seasonally in humans. It also found 2 compounds (Homocysteine and Glutaric acid) that are not known to vary seasonally and one that varies seasonally in animals (Corticosterone) but with no human studies currently. I focused on the 22 SVBs that are confirmed to vary seasonally in humans by published studies indexed by PubMed. I used these as input for the enrichment algorithm. **Table 12** contains the references supporting the seasonal relationship and references that refute the relationship, if any exist.

I combined results from my SeaWAS study with a carefully curated set of diseases related to birth month that I developed previously. This file is available with the supplement and includes the PubMed ID, publication year, disease area (high-level disease category), and a binary variable indicating whether the study found or failed to find the association. In this feasibility study of the algorithm's framework, 7 randomly chosen BMDDs were selected with one constraint: that the number of distinct disease genes differed largely. For example, 21 genes were implicated in reproductive performance while 1253 genes were implicated in asthma.

SVBs and BMDDs were mapped to DisGeNET to extract genes associated with each SVB and BMDD. Examples of the extraction process are given in methods **Table 9** and **Table 10.** For this study, I ran the DisGeNET extraction on 7 random BMDDs with different gene set sizes. I also used DisGeNET to extract the genes related to the 22 SVBs given in **Table 12** (folate and folic acid are counted as separate SVBs but merged under vitamin B9 in **Table 12).**

5.4.2 Enrichment Results

The enrichment algorithm investigates the overlap among gene sets from the BMDD and each SVB. It compares this overlap to an average across 100 randomly generated gene sets of the same size (i.e., number of genes) as the particular BMDD of interest. The average overlap score from the 100 random sets is compared against the actual number to determine significance using Fisher's exact test. P-values were adjusted using the Bonferroni correction method. I then ranked each significant BMDD-SVB pair by the ORs. Results are shown in **Table 13** with the top three associations in bold.

The top SVBs associated with each BMDD are biologically intuitive. Cardiovascular disease is known to involve vitamin K regulation with the anti-coagulant drug warfarin targeting the well-studied vitamin K gene: VKORC1. The two top SVBs related to asthma (a known immune-related condition) are also immune related: eosinophils and neutrophils (Kidd et al., 2016). Atrial fibrillation's top hits are calcium related and atrial fibrillation is associated with increased calcium release from the sarcoplasmic reticulum (Hove-Madsen et al., 2004).

5.4.3 Restrict to Genes Involved in Developmental Processes

Genes were restricted to include only genes involved in at least one developmental process using GO annotations. This was primarily because genes that are involved in developmental processes

are more likely to play a role in birth month associations. This reduced the number of potential genes as shown in **Table 14.**

Table 14 illustrates how the algorithm started with 1,253 genes associated with Asthma as extracted from DisGeNET. Because asthma is also known to be associated with birth month and a BMDD, I ran the algorithm to find overlapping genes between asthma and SVBs where the overlapping genes were enriched. This reduced the number of genes potentially involved in a seasonally varying process at birth down to 439 genes from 1253. I then restricted these 439 genes to only include genes known to be involved in some developmental process using GO term annotations. This further reduced the number of genes down to 140. Therefore, only 11.2% of asthma-related genes are potentially involved in developmental processes related to SVBs that could potentially lead to birth month-related effects.

5.4.4 Developmentally Expressed Genes Link SVBs to BMDDs in Biological Graphs

My algorithm produces output containing each BMDD, the enriched SVBs and the overlapping genes that are developmentally expressed between the BMDD and the SVB. A file containing tuples of BMDD, SVB, and gene is available in the supplement.

For illustrative purposes, I show three SVB graphs from this feasibility study: two immune cells (eosinophil and neutrophil) and one hormone (parathyroid hormone). Full resolution images are available with supplemental information. The immune cells are shown in **Figure 23,** the SVB is represented by a triangle, the BMDD is represented by a square and circles represent the overlapping genes. Cluster annotations from DAVID are shown above or near each cluster. In the neutrophil graph (**Figure 23A**) there are clusters involving blood vessel development, response to an organic substance, positive regulation of nitrogen compound metabolic processes, regulation of cell proliferation and embryonic development / birthing. In **Figure 23B,** the largest

cluster includes genes related to the immune response (as expected). Other clusters are for neuron development, tube development (e.g., neural, endothelial tubes), response to sterol hormone synthesis and insulin stimulus. The same five BMDDs are involved in both: attention deficit hyperactivity disorder (ADHD), cardiomyopathy, atrial fibrillation, cardiovascular disease, and asthma. Eosinophils and neutrophils are in the top three most enriched SVBs for both asthma and ADHD (**Table 13**). Its possible that the genes involved in neuron development is responsible for the ADHD – Eosinophil relationship (**Figure 23B).**

Figure 24 shows the graph for parathyroid hormone (PTH). There are three main genetic processes involved: positive regulation of the development process, receptor linked signal transduction, and response to hormone stimulus. Not only is parathyroid hormone a hormone, but it also is involved in regulation or other hormones. Both PTH and vitamin D are hormones that regulate each other through complex mechanisms. Positive regulation of the development process is also enriched in this graph connecting PTH and BMDD this fits with the involvement of this SVB with a contribution to disease risk that is due to birth month.

Table 12. Biofactors With Seasonal Dependencies Extracted from the Literature

		Seasonal Relationship	
Biofactor	**Notes**	**Reference Supporting**	**Reference Refuting**
Hormone			
Parathyroid Hormone (PTH)	PTH and vitamin D are slightly anti-correlated	(Meier et al., 2004; Steingrimsdottir et al., 2005)	
Estrogen	(modulated through Vitamin D)	(Lee et al., 2012)	
Estradiol		(Bjørnerem et al., 2006)	
Testosterone	(modulated through Vitamin D)	(Lee et al., 2012)	
Progesterone	(modulated through Vitamin D)	(Lee et al., 2012)	
Vitamins/Minerals			
Vitamin A (retinol, beta-carotine)	Bitot eye spots are a sign of vitamin a deficiency	(Basu et al., 1994; Khan and Khan, 2005; Woodhouse and Khaw, 2000; Xiang et al., 2008)	
Vitamin B9: Folate and Folic Acid		(Hao et al., 2003; McKinley et al., 2001)	
Vitamin B12		(Palva and Salokannel, 1972)	
Vitamin C (Ascorbic acid)		(Hallmann et al., 2013; Paalanen et al., 2013; Woodhouse and Khaw, 2000)	
Vitamin D		(Douglas, 1993; Meier et al., 2004; Steingrimsdottir et al., 2005)	
Vitamin E		(Woodhouse and Khaw, 2000)	
Vitamin K	Vitamins K and D regulate osteocalcin	(Anai et al., 1991; Douglas, 1993)	
Calcium		(Douglas, 1993; Meier et al., 2004; Steingrimsdottir et al., 2005)	
Phosphate		(Douglas, 1993)	
Immune Cells			
Neutrophil		(Gelardi et al., 2014; Klink et al., 2012)	
Eosinophil		(Gelardi et al., 2014; Henriksen, 1986; Liu, 1988)	
Basophil		(Liu, 1988)	(Henriksen, 1986)
Other Cells/Metabolites			
Hemoglobin		(Lee et al., 1987)	
Uric	Uric acid	(Parks et al., 2003)	
Creatine		(Percy et al., 1982)	
Lactic	Lactic acid	(Svedenhag and Sjödin, 1985)	

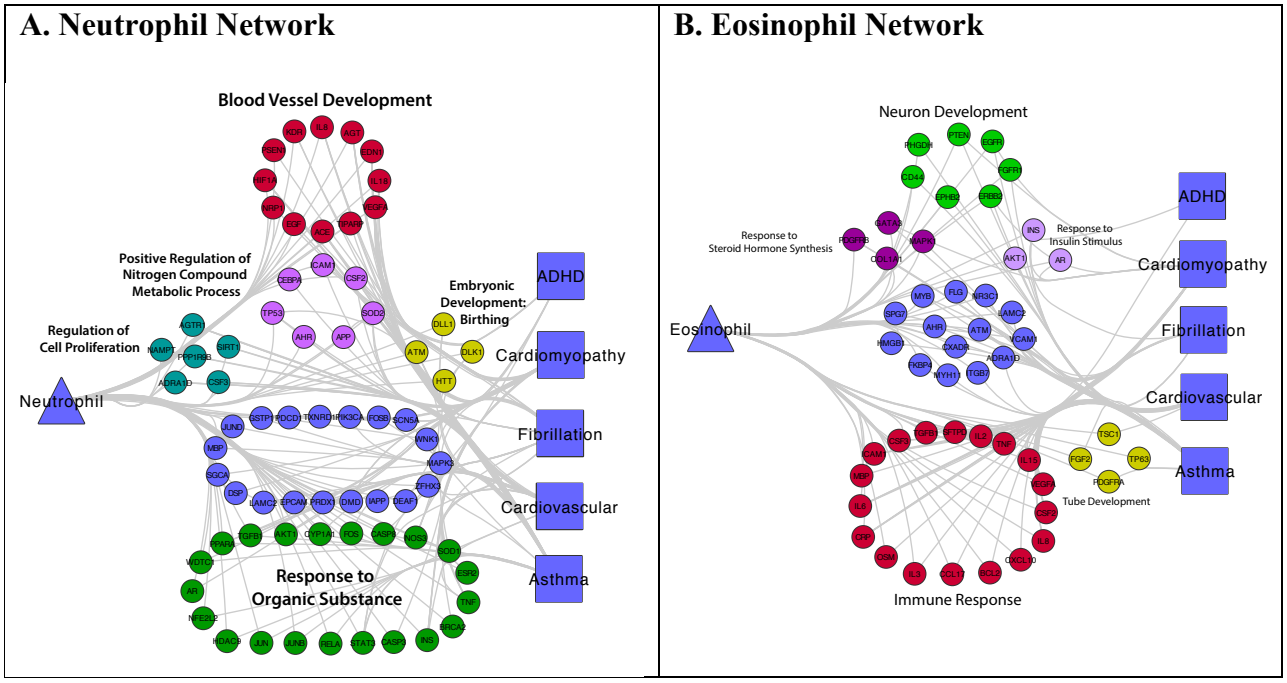


Figure 23. Immune Cell Bi-Partite Networks Connecting SVBs to BMDDs Via Overlapping Genes Involved in Developmental Processes. Full resolution images are available on figshare (link in methods).

Table 13. BMDD-SVB Enriched Overlapping Gene Sets Sorted by OR

BMDD Disease	Enriched SVB	OR	-log(p) *
Asthma	**Eosinophil**	**23.545308**	**92.3610942**
	Neutrophil	**7.608800**	**41.7971736**
	Vitamin D	**7.597455**	**4.9508901**
	Phosphate Gene Set 1 (Phosph)	6.787030	36.4156313
	Hemoglobin	6.684969	19.1382008
	Uric	5.672420	12.1008132
	Calcium Gene Set 2 (Calcinosis)	4.085383	5.4623019
	Calcium Gene Set 1 (Calci)	3.511367	9.7828493
	Phosphate Gene Set 2 (Phosphate)	5.095905	10.1171083
	Parathyroid Hormone	4.494405	23.8948578
ADHD	**Eosinophil**	**3.805091**	**5.2892789**
	Parathyroid Hormone	**3.553309**	**11.9138662**
	Neutrophil	**3.308624**	**9.2041648**
	Phosphate Gene Set 1 (Phosph)	3.302909	9.7728579
	Calcium Gene Set 1 (Calci)	2.679110	7.5585981
	Calcium Gene Set 2 (Calcinosis)	2.043983	7.1930726
Fibrillation	**Calcium Gene Set 2 (Calcinosis)**	**12.287454**	**7.1930726**
	Phosphate Gene Set 1 (Phosph)	**6.440454**	**9.7728579**
	Parathyroid Hormone	**6.437664**	**11.9138662**
	Calcium Gene Set 1 (Calci)	6.380333	7.5585981
	Neutrophil	6.252987	9.2041648
	Eosinophil	6.242087	5.2892789
Reproductive	**-**	-	-
Cardiovascular	**Vitamin K**	**45.187116**	**5.8861829**
	Folic Acid	**14.548556**	**6.5377502**
	Vitamin D	**13.319860**	**6.1949509**
	Phosphate Gene Set 2 (Phosphate)	11.560768	23.1079644
	Phosphate Gene Set 1 (Phosph)	9.246976	39.8563236
	Uric	9.174204	17.0939677
	Calcium Gene Set 2 (Calcinosis)	8.543921	13.0824611
	Calcium Gene Set 1 (Calci)	8.492730	28.6820297
	Neutrophil	7.892360	28.6839072
	Eosinophil	6.946996	19.5349288
	Hemoglobin	6.384460	11.7733926
	Parathyroid Hormone	4.149517	13.8335245
Cardiomyopathy	**Vitamin D**	**10.145258**	**3.851662**
	Eosinophil	**8.923802**	**25.690966**
	Lactic	**8.159469**	**6.974233**
	Hemoglobin	7.666338	15.743251
	Calcium Gene Set 2 (Calcinosis)	7.616024	10.981833
	Neutrophil	6.468582	21.428566
	Phosphate Gene Set 1 (Phosph)	6.284259	22.014083
	Calcium Gene Set 1 (Calci)	6.049040	16.056148
	Uric	5.784556	6.909913
	Parathyroid Hormone	4.781197	16.279232
	Phosphate Gene Set 2 (Phosphate)	4.342993	3.422636
Mitral Valve	**Phosphate Gene Set 1 (Phosph)**	**16.780523**	**5.1995428**
	Calcium Gene Set 1 (Calci)	**13.587347**	**3.2560223**

* greater than 3.0 is significant after Bonferroni correction

Table 14. Number of Genes Involved in BMDD That Are Potentially Involved in Birth Month Contribution to Disease is Drastically Reduced After Framework Is Applied

BMDD	No. Distinct Genes (A)	No. of Overlapping Genes from Enriched BMDD-SVB Pairs (B)	No. of Genes from B Involved in Developmental Processes (C)	% of Genes Potentially Involved in Birth Month Contribution Out of All BMDD Genes (C / A)
Asthma	1253	439	140	0.112
ADHD	338	63	18	0.053
Fibrillation	318	105	45	0.142
Reproductive	21	-	-	-
Cardiovascular	775	302	109	0.141
Cardiomyopathy	717	250	89	0.124
Mitral Valve	82	24	15	0.183

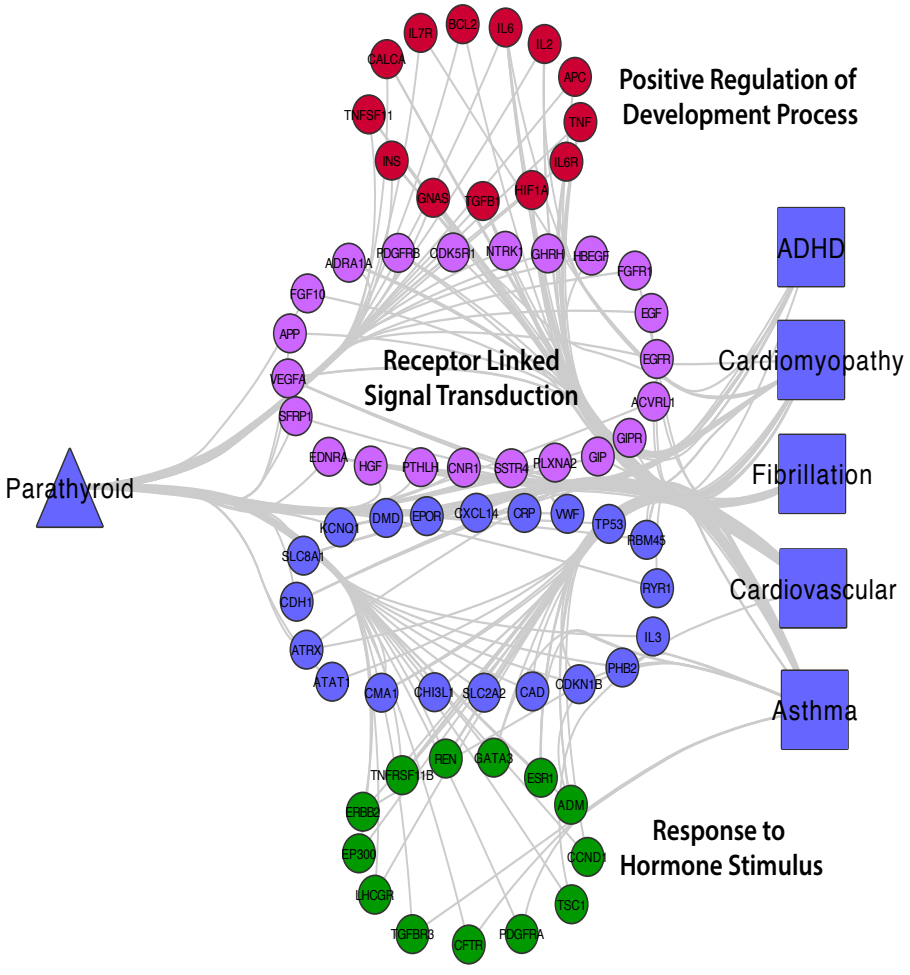


Figure 24. Graph Connecting Parathyroid Hormone with BMDDs Via Overlapping Genes Involved in Developmental Processes. Full resolution images are available on figshare.

5.5 Discussion

5.5.1 Value of a High-throughput Birth Month-Disease Dependency Genetic Algorithm

The relationship between genes, environment and disease has been discussed by medical researchers since the early days of genetics (Boland et al., 2013b). Because the relationship between genes and the environment is complex, researchers originally investigated single environmental exposures and how those exposures influenced disease risk via genetic changes (Wei et al., 2012). Later the Environment-Wide Association Study (EWAS) was developed, which explored a large variety of environmental exposures (not just one as was done previously) and then explored how those exposures effected one single disease: Type 2 Diabetes (Patel et al., 2010). While an improvement over previous work, EWAS was still limited to exploring one disease at a time.

The original SeaWAS study, described in chapter two, revealed multiple birth month-disease dependencies (called BMDDs). Additionally, 92 other articles revealed additional information on BMDDs. None of these epidemiological studies sheds light on the genetic underpinnings of BMDDs. Therefore, a method was required that could investigate diverse environmental triggers across a plethora of diseases and disease types. To address this gap, I developed an algorithmic framework to uncover enriched SVBs related to BMDDs described in this chapter.

In addition to finding SVBs enriched in BMDDs, I also explored the overlapping genes implicated in both the SVB and the BMDD. I limited my investigation to only those genes that are known to be involved in developmental processes to hone in on genes that are potentially responsible for birth month disease dependencies.

5.5.2 Highlighting One Well-Studied Disease: Asthma

The top SVBs enriched for asthma were eosinophil (OR=23.545), neutrophil (OR=7.601) and vitamin D (OR=7.597) (**Table 13**). The relationship between eosinophils (key cells in the immune response) and asthma is well known and studied (Busse and Sedgwick, 1992). At the same time asthma exacerbations and underlying gene expression changes are also known to vary seasonally (Bjornsdottir et al., 2011). As revealed by my framework, there is also literature support for a relationship between eosinophil changes and season (Gelardi et al., 2014; Henriksen, 1986; Liu, 1988). My framework revealed 42 genes in common between asthma and eosinophils that are also involved in developmental processes (the entire eosinophil network is shown in **Figure 23B)**. The immune response was the key functional process involved along with neuron development (which could help to explain the interesting relationship between eosinophils and ADHD).

Table 13 reveals an interesting relationship between asthma and ADHD: they both share eosinophils and neutrophils among their top three SVB enrichments (asthma also has vitamin D and ADHD has parathyroid hormone—which are also related to each other). When I investigated the biological network (**Figure 23B**), the enrichment in neuron development genes is revealed. Importantly, asthma patients are known to be at increased risk for developing ADHD and this increased risk was observed even after adjusting for urbanization and comorbid allergic diseases suggesting an underlying etiology behind the two diseases (Chen et al., 2013). Others have also studied the relationship between asthma and ADHD (Secnik et al., 2005) without uncovering a clear genetic/biological mechanism for the relationship. Importantly, my framework enables researchers to construct biological networks that connect complex associations between SVBs and BMDDs through their shared underlying genetic pathways. This enables researchers to formulate and test hypotheses behind disease etiology and progression.

5.6 Limitations

This study is limited by the information contained and available on PubMed regarding biofactors that vary seasonally (SVBs) and BMDDs. Therefore, neither of these lists is fully complete as there may be other studies not reported in PubMed, studies not translated into English and so forth that would prevent utilization of their findings in my framework. My initial query to PubMed returned 3,627 articles related to seasonal variation in humans. Because I was primarily interested in biological compounds that vary seasonally (such as hormones, vitamins, immune cells), I added additional query terms (as specified in the methods section). Manually reviewing all 3,627 articles was not feasible; therefore some lesser-known SVBs may be missing. Additionally, the fetal-maternal barrier warrants further investigation as the placenta is known to be susceptible to environmental effects (Nelissen et al., 2011). Incorporating knowledge on the epigenetics of the placenta could help with understanding the underlying disease mechanism (Nelissen et al., 2011).

5.7 Conclusion

In this chapter, I present a framework that combines existing data repositories (PubMed, GO, and DisGeNET) to uncover biological mechanisms underlying birth month – disease dependencies (BMDDs) using known Seasonally Varying Biofactors (SVBs). This framework allows me to link epidemiological data on birth month-disease relationships and genetic data on gene-disease associations recorded in existing public data repositories. My algorithm finds enriched BMDD-SVB pairs using the genes involved in both the disease and the SVB. I then investigate the overlapping genes in these enrichments and trim away genes not known to be involved in developmental processes using GO annotations. My framework produces a bipartite graph that connects enriched SVBs with BMDDs through their overlapping developmental gene sets. Thus

allowing the formation of biological hypotheses around the genetic mechanisms underlying birth month-disease dependencies. As a proof-of-concept, results are presented from 7 BMDDs across all identified known SVBs.

5.8 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in the American Medical Informatics Association's Translational Science Proceedings (Boland and Tatonetti, 2016c). Support for this research provided by R01 GM107145 (MRB, NPT). MRB was supported by the National Library of Medicine training grant T15 LM00707 (MRB) from Jul 2014 – Jun. 2016 when this work was conducted.

Chapter 6

A Phenocopy of the Birth Season – Disease Effect: 7-DehydroCholesterol Reductase

6.1 Abstract

Seasonal factors occurring during the prenatal or perinatal period can affect long-term disease outcomes. In chapter two, I described the use of clinical data from 1.7 million patients to identify 55 diseases correlated with birth season. However, this work did not reveal the environmental drivers underlying these associations. Vitamin D is a seasonally varying compound synthesized in the skin using 7-dehydrocholesterol and ultraviolet B radiation. A competing reaction occurs to convert 7-dehydrocholesterol to cholesterol via 7-dehydrocholesterol reductase (DHCR7). Deleterious mutations in *DHCR7* decrease the ability to produce cholesterol and enhance production of vitamin D. Thus making *DHCR7* mutations more evolutionarily favored in regions of the world with historically low sunlight access.

Importantly, a rare Mendelian disease - Smith-Lemli-Opitz Syndrome (SLOS [MIM 270400]) is characterized by compound heterozygous mutations in *DHCR7*. SLOS results in severe fetal

deformities and malformations. In this study, I discuss the toxicological information gleaned from a deep exploration of *DHCR7*. First, I present a compilation of SLOS-inducing *DHCR7* mutations and the geographic distribution (S. America, Europe, Australia, Asia) of those mutations among diseased populations. I describe several hypotheses for *DHCR7* mutations that would maximize vitamin D production via increased evolutionary pressure. Next, I looked at the mutational spectrum of *DHCR7* in an ethnically diverse, presumed healthy population from ExAC (http://exac.broadinstitute.org/). This study observed that several mutations thought to be disease causing occur in healthy populations as well, sometimes with high frequencies, indicating an incomplete understanding of SLOS and highlighting new research opportunities. I also highlight several *DHCR7* variants found in ExAC that could represent SLOS carriers in under-represented ethnic groups.

Knowledge of the importance of vitamin D during the prenatal period coupled with an enhanced global understanding of the *DHCR7* mutational spectrum allowed me to hypothesize that exposure to DHCR7 inhibitors would result in deleterious effects similar to SLOS. I tested this hypothesis by investigating the fetal outcomes following prenatal exposure to DHCR7 modulators. First-trimester exposure to DHCR7 inhibitors resulted in outcomes similar to those of known teratogens (50% vs. 48% born-healthy). DHCR7 activity should be considered during drug development and prenatal toxicity assessment.

6.2 Introduction

Mendelian diseases are genetic conditions that follow a 'traditional' pattern of inheritance. Previously, researchers utilized information from Mendelian gene mutations to study shared underlying disease mechanisms that are common to non-Mendelian diseases in complex diseases (Blair et al., 2013) and cancer (Melamed et al., 2015). Mendelian diseases are also useful in

studying developmental effects of gene mutations and can help researchers understand the effects of a potential pharmaceutical target or off-target effect (Bunnage et al., 2015) increasing the impact of their discoveries (Fishman and Porter, 2005). Understanding the underlying mechanisms of Mendelian diseases can enable *a priori* prediction of fetal outcomes following prenatal pharmaceutical exposure.

In this review, I detail one orphan Mendelian disease - Smith-Lemli-Opitz Syndrome (SLOS) resulting from mutations in 7-dehydrocholesterol reductase (DHCR7). These mutations affect a pathway involving vitamin D and cholesterol production. Mutations affecting vitamin metabolism can play an important role in drug response (Carr et al., 2009). In-depth study of this biological pathway enables us to explain off-target effects of prenatal drug exposure and highlights DHCR7's importance in drug development for potential prenatal toxicity assessment.

6.2.1 Clinical Characteristics

SLOS was first identified in 1964 when physicians described a similar pattern of congenital anomalies, including mental retardation, incomplete external genitalia, and abnormalities of face, hands, and feet that followed a familial inheritance pattern (Smith et al., 1964). Later it was discovered that extremely high 7-dehydrocholesterol levels and surprisingly low serum cholesterol levels were common biomarkers of SLOS. This led to the discovery of the exact location in the cholesterol synthesis pathway that was defective in SLOS patients, namely the conversion of 7-dehyrocholesterol into cholesterol (the last step in cholesterol biosynthesis) (Tint et al., 1994). Subsequently, DHCR7 was identified as the culprit gene (Fitzky et al., 1998). DHCR7 is the only enzyme that converts 7-dehydrocholesterol to cholesterol (Wilcox et al., 2007). Cholesterol cannot be produced without DHCR7.

The physical presentation of SLOS differs widely among individuals, varying by severity,

genotype, and other environmental factors (Jira et al., 2003). The most frequently occurring feature is 2/3 toe syndactyly (i.e., "webbed toes") occurring among 97% of patients followed by mental retardation with 95% of patients (Jira et al., 2003; Kelley and Hennekam, 2000). Other common signs include microcephaly (84%), postnatal growth retardation (82%), anteverted nares (78%), ptosis (70%), genital anomalies (65%), and congenital heart defects (among 54% of SLOS patients) (Jira et al., 2003; Kelley and Hennekam, 2000). SLOS severity ranges across a wide spectrum. Some SLOS patients present with a mild form (Prasad et al., 2002) with minimal symptoms and no developmental delay (Nowaczyk and Irons, 2012). Others have a severe form that can result in a lack of sexual dismorphism with a functional XY karyotype and female internal and external genitalia (Fukazawa et al., 1992).

The importance of cholesterol in prenatal embryonic and fetal development, and its partial to complete absence in SLOS, helps to explain the pleotropic phenotypes within SLOS. In patients possessing homozygous null mutations in DHCR7, cholesterol production is absent and prenatal lethality results (Lanthaler et al., 2013). Other mutations reduce DHCR7 expression to less than 5%, dramatically reducing cholesterol production in the body (Fitzky et al., 1998).

6.2.2 Genetic Characteristics

SLOS is an inherited autosomal recessive disease with each parent contributing one mutated copy of DHCR7. Inheritance follows a compound heterozygosis pattern whereby each parent contributes one copy of ***different*** mutations in DHCR7. Therefore, the SLOS patient is heterozygous for two mutations. Being heterozygous for only one mutation generally does not cause the SLOS phenotype, although instances have been reported (De Brasi et al., 1999; Fitzky et al., 1998). Being homozygous for a null mutation in DHCR7 typically results in prenatal death (Lanthaler et al., 2013). This explains why most full-term viable SLOS patients are compound

heterozygotes. **Figure 25** depicts the autosomal inheritance of SLOS in children and how compound heterozygosity is responsible for the disease phenotype. The discrepancy between the DHCR7 mutation carrier rate and SLOS incidence (Nowaczyk et al., 2006) is believed to result from prenatal loss of individuals with homozygous null mutations during the first trimester (Lanthaler et al., 2013). As in many inherited genetic conditions, *de novo* mutations have also been reported (Waye et al., 2007).

Importantly, the relationship between null mutations in DHCR7 and SLOS severity is not one-to-one because variations in the maternal genome can increase the amount of cholesterol passed by the placenta to the fetus (Lanthaler et al., 2013) modulating the offspring's phenotype. Because cholesterol is critical during early development, having increased prenatal cholesterol levels distributed from the mother via the placenta can mitigate many SLOS symptoms (Lanthaler et al., 2013). Diverse factors modulate SLOS severity, therefore it is not possible to completely predict the disease phenotype using genotype information alone or vice-versa (Witsch-Baumgartner et al., 2013).

6.3 Compendium Containing SLOS-Inducing DHCR7 Mutations

6.3.1 Development of the DHCR7 SLOS Mutation Compendium

SLOS patients are compound heterozygotes for diverse mutations in DHCR7 or homozygous for non-null mutations. SLOS is a rare disease (~1 in 40,000) and studies typically involve only small cohorts of patients. Even with small cohorts, many DHCR7 mutations have been reported. Therefore, I reviewed the literature, and developed a compendium of SLOS-inducing DHCR7 mutations. I collected the reported frequencies for each mutation across all studies, and, when available, I extracted the patient genotype (i.e., compound heterozygous mutant alleles) and their geographic location or ethnicity. Specifically, I am interested in the ethnicities of SLOS patients

and their corresponding genotypes because certain gene variants affect outcomes only within a given ethnicity (e.g., ACBCB1 variant in Caucasians) (Megías-Vericat et al., 2015).

To develop the compendium, I analyzed all publically available DHCR7 mutations contained in the Human Gene Mutation Database (HGMD) (HGMD, 2015; Stenson et al., 2014) and their corresponding publications (HGMD last accessed in May 2015). I kept track of patients' DHCR7 genotype information and their ethnicity/geographic origin when available (described later). The aggregated allele frequencies for all DHCR7 mutations across 30 studies are given in **Table S1** of the published paper (Boland and Tatonetti, 2016a). This review focuses only on deleterious mutations to DHCR7, and therefore does not include 13 silent DHCR7 mutations (Waterham and Hennekam, 2012). Using the HGMD, I found 138 distinct publically available mutations to DHCR7 and 165 reported (with the additional 27 mutations being proprietary, and only accessible via paid membership).

This literature review revealed additional DHCR7 mutations (by investigating papers cited in those retrieved papers) and overall my compendium contains 147 DHCR7 mutations, 145 of these being SLOS-inducing mutations. One mis-sense mutation, W158C, was found in an unaffected sibling of a SLOS individual (noted in **Table S1),** and another mutation, G344D, was reported in a patient with holoprosencephaly (noted in **Table S1)** a different mutation at the same position, G344R, was reported in a SLOS patient. Mutations in DHCR7 that result in holoprosencephaly were also included in the compendium and denoted in **Table S1.**

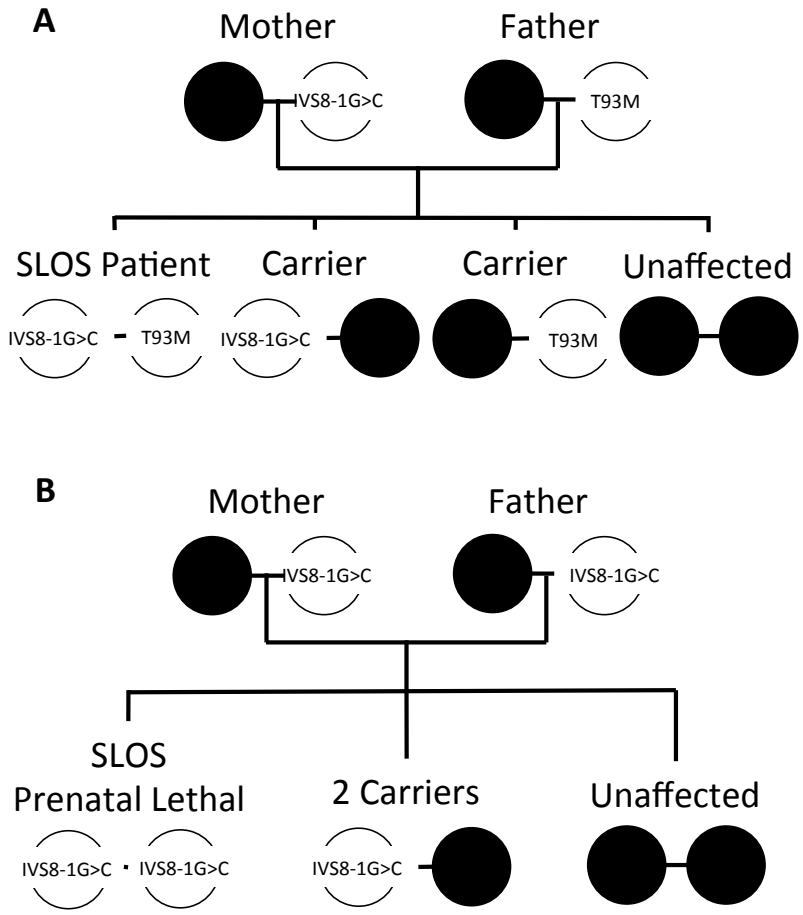


Figure 25. Full-Term Smith-Lemli-Opitz Syndrome (SLOS) Patients Are Typically Compound Heterozygous for Two Distinct Mutations in DHCR7 (Figure 25A) while Homozygous Null Individuals are Detected Less Frequently Due to Prenatal Lethality (Figure 25B).

Figure 25 depicts the autosomal inheritance of SLOS in children and how compound heterozygosity is responsible for the disease phenotype. Many SLOS genetic studies focus on compound heterozygous patients (**Figure 25A**) because most homozygous phenotypes result in prenatal fatalities, reducing the detection rate (**Figure 25B).** Both W151X and IVS8-1G>C are null mutations in DHCR7 meaning that they reduce DHCR7 expression to almost 0% in the homozygous state. Therefore if an individual is homozygous for either of these mutations or heterozygous for the combo then little to no DHCR7 expression would result (Correa-Cerro et al., 2005). On the other hand, T93M is a non-null mutation in DHCR7 that reduces DHCR7 expression by 5% when compared to normal (Nowaczyk et al., 2004a). Therefore a compound heterozygous patient with one IVS8-1G>C null mutation and one T93M mutation would have around 45% functional DHCR7 and SLOS would result, but prenatal fatality would be averted **(Figure 25A)**.

Table 15. DHCR7 Mutations Implicated in SLOS with Allele Frequency >=1% Across 30 Studies

Allele's Effect on Coding Sequence	Genomic Chromosome Position[†]	Accession Number (RS ID) [†]	Intron/Exon	Allele Freq. in 523 SLOS patients, N=1,037 alleles (%)	Allele Freq. in ExAC Population, M=60,706 healthy individuals (%)
IVS8-1G>C*	71146886	rs138659167	Intron 8	291 (28.4)	386 (4.2×10^{-1})
T93M	71155082	rs80338853	Exon 4	96 (9.4)	**3** (2.7×10^{-3})
W151X	71152447	rs11555217	Exon 6	86 (8.4)	82 (6.8×10^{-2})
V326L[††]	71146873	rs80338859	Exon 9	52 (5.1)	5 (4.8×10^{-3})
R404C	71146639	rs61757582	Exon 9	36 (3.5)	4 (3.5×10^{-3})
R352W[††]	71146795	rs80338860	Exon 9	34 (3.3)	2 (1.7×10^{-3})
E448K	71146507	rs80338864	Exon 9	23 (2.2)	1 (8.5×10^{-4})
R352Q	71146794	rs121909768	Exon 9	22 (2.2)	4 (3.4×10^{-3})
G410S	71146621	rs80338862	Exon 9	15 (1.5)	5 (4.3×10^{-3})
Unidentified Suspected Variant	-	-	-	13 (1.3)	-
P51S	-	-	Exon 4	12 (1.2)	-
R242C	71150032	rs80338856	Exon 7	12 (1.2)	**10** (8.3×10^{-3})
F302L	71148915	rs80338858	Exon 8	12 (1.2)	1 (8.3×10^{-4})

[†] Obtained from ExAC output, accessed in November 2015 (http://exac.broadinstitute.org/)
*Annotated as: c.964-1G>C in some literature articles
[††] SLOS patients with single mutations (true heterozygotes) were found (3 patients), 2 had mutations in V326L and 1 had a mutation in R352W (Fitzky et al., 1998). Both mutations were shown to reduce expression of DHCR7 upon heterologous expression by >90% (Fitzky et al., 1998).

Table 16. Top 10 DHCR7-SLOS Inducing Mutations Ranked By Frequency in ExAC Population

	ExAC Allele Frequency (%)								
Allele's Effect on Coding Sequence	**Overall**	**African**	**East Asian**	**European (Non-Finnish)**	**Finnish**	**Latino**	**Other**	**South Asian**	**Overall Frequency (%) Among SLOS Patients**
IVS8-1G>C*	0.420	0.295	0	**0.677**	0.177	0.165	0.149	0.007	28.062
W151X	0.068	0.029	0	**0.116**	0	0.009	0.111	0	8.39
V330M	0.036	0.012	0.012	0.044	0.018	**0.082**	0	0.007	0.096
T154R	0.008	0.010	0	**0.014**	0	0	0	0	0.193
R242C	0.008	**0.029**	0.023	0.008	0	0	0	0	1.157
S169L	0.008	0.010	0	0.012	0	0	**0.110**	0	0.868
G303R	0.007	0.010	**0.046**	0.005	0	0.009	0	0	0.579
R363C	0.007	0	0	0.011	**0.015**	0	0	0	0.096
G147D	0.007	0	0	**0.012**	0	0	0	0	0.675
L157P	0.007	0	0	**0.012**	0	0	0	0	0.675

Ethnicity with Highest Allele Frequency For Each DHCR7 SLOS-Inducing Mutation is **Bolded**

6.3.2 Common SLOS-Inducing DHCR7 Mutations

I found that 12 mutations (of 145 SLOS-inducing mutations) occurred with allele frequency of at least 1% across the **30** reviewed studies (**Table 15**). The most common of these were IVS8-1G>C (28.446%), T93M (9.384%), W151X (8.407%) V326L (5.083%), R404C (3.519%), and R352W (3.324%). I compared this result to data obtained from 60,706 assumed healthy individuals (i.e., non-SLOS) to determine the frequency of various DHCR7 mutations in that reference population. I used the ExAC database (Lek et al., 2015) available at http://exac.broadinstitute.org/ (accessed November 2015). These results are also given in **Table 15.** The majority of frequent DHCR7 mutations in SLOS patients also occur at higher frequencies in the ExAC population (this is intuitive). There are two exceptions: T93M that was common among SLOS patients (9.4%) but rare in the healthy population (only 3 allele mutations observed out of 60,706 individuals). Another counter-intuitive result was that R242C was found relatively frequently in the ExAC population (10 allele mutations observed), but was only mutated in 1.2% for SLOS. I present the top 10 DHCR7 SLOS-inducing mutations as ranked by their incidence in the ExAC population and their corresponding incidence in the SLOS patient set in **Table 16.** All of the SLOS DHCR7 mutations were less than 1% frequency in the ExAC population. I compared the entire compendium of DHCR7 mutations to ExAC. The results of this comparison including accession numbers and chromosomal locations are given in **Table S2** (of the published paper).

A SLOS review from 2005, mentioned 11 mutations with at least 1% frequency across SLOS patients (Yu and Patel, 2005). I compared my findings to theirs and noticed that the top 8 mutations remain unchanged. I noted that two additional mutations were on the list, namely R352Q (2.151%) and P51S (1.173%) and one mutation on their list was not included in mine,

namely S169L (Yu: 1.6%) having a frequency of 0.868% in my compilation of 30 studies (**Table S1, Table 15).**

The availability of Korean genetic data resulted in the addition of R352Q (Oh et al., 2014) to the list of frequent SLOS mutations (>=1%). Importantly, while SLOS occurs more frequently among Europeans (Nowaczyk et al., 2001; Nowaczyk et al., 2006)**,** mutations in DHCR7 exhibit an evolutionary pressure towards Northern Europeans and Northeast Asians (Kuan et al., 2013), indicating that SLOS may be currently under-reported among Northeast Asians. Therefore, as more genotype data becomes available from non-European SLOS patients, I would expect the common allele frequencies to change somewhat and become less biased towards Northern Europeans.

A couple of DHCR7 mutations were reported to cause SLOS in the heterozygous form (De Brasi et al., 1999; Fitzky et al., 1998; Patrono et al., 2002; Waye et al., 2005). However, it is likely that an unidentified suspected variant was present in these instances, as SLOS results from two defective copies of DHCR7, either in a compound heterozygous state (2 different DHCR7 mutations) or a homozygous mutated state. Importantly some asymptomatic siblings of SLOS patients were heterozygous carriers of the null W151X mutation (they should have around 50% expression of DHCR7) and they exhibited no symptoms of SLOS suggesting that the disease causing state requires DHCR7 expression to be <50% (Fitzky et al., 1998). This also confirms that one copy of a DHCR7 mutation does not result in SLOS.

6.4 DHCR7 Mutations Vary By Geographical Location

6.4.1 Specific DHCR7 Mutations Exhibit Geographic Dependency

SLOS incidence ranges by geographical location and ethnicity because carrier frequencies vary

by region (Nowaczyk et al., 2006). Reports range from 1 in 9,000 from Czechoslovakia and Central European populations to 1 in 40,000-70,000 in Canada (Nowaczyk et al., 2006; Nowaczyk et al., 2004b). The incidence in both the United States of America (USA) and Europeans is 1 in 30,000 live births (Nowaczyk et al., 2001; Nowaczyk et al., 2006). Variations in reported incidence rates are thought to be due to sampling bias differences. The carrier rate, or the proportion of individuals with one copy of a known SLOS-inducing DHCR7 mutation, is estimated to be 3% among persons of European ancestry (Nowaczyk et al., 2001) while another study found 1-2% among Caucasians (Porter, 2008). In the USA the carrier frequency is 1% (Cross et al., 2014; Waye et al., 2002), although in the state of Utah the carrier rate is 4% (Opitz et al., 2002).

Mutation frequencies are known to vary by geographic location / ethnicity (Fricke-Galindo et al., 2016). Likewise, the spectrum of DHCR7 mutations varies by geographical location. There is evidence that T93M is the founder mutation with origins in the Mediterranean basin (Nowaczyk et al., 2004a) and is common in Italy, Spain, (Kozák et al., 2000; Witsch-Baumgartner et al., 2001; Witsch-Baumgartner et al., 2005) and Portugal (Cardoso et al., 2005). On the other hand, Northern Europeans (Austrians, Germans, Dutch) present with the W151X mutation more frequently (Lanthaler et al., 2013; Witsch-Baumgartner et al., 2005). In Korea, three mutations account for most of the observed variants (Oh et al., 2014). Persons of African ancestry rarely develop SLOS, however some carriers were observed among persons of African ancestry (Nowaczyk et al., 2001). One SLOS patient with Spanish-African mixed ancestry was identified, this patient had T93M (the common Spanish SLOS allele) and V281M (Nowaczyk et al., 2004a). The carrier frequency of DHCR7 mutations among African Canadians was low at 0.79% (Waye et al., 2002).

The relationship between the specific combinations of DHCR7 mutations and ethnicity/geographic location of SLOS patients warranted further exploration. Therefore, I extracted all genotype data for SLOS patients that also contained ethnicity details, or where the ethnicity could be deduced (e.g., a German study enrolling patients in Germany). Not all studies listed in **Table S1** reported the genotype and ethnicity information for DHCR7 mutations. The aggregated results are shown in **Table 17** with genotype data from 21 studies representing 229 patients from 26 different countries/ethnicities (if mixed ethnicities are included). The exon distribution of the 1024 alleles with DHCR7 mutation information available is shown in **Figure 26A.** The proportion of DHCR7 mutations that exist in a trans-membrane domain are shown in **Figure 26B.** Three protein structure models are shown in **Figure 26B** because the 3-dimensional structure of DHCR7 has not been solved yet. The three models are: the human protein reference database (HPRD) containing trans-membrane exon bound information, the Waterham *et al.* study (Waterham and Wanders, 2000), and the Fitky *et al.* study (Fitzky et al., 1998). I also analyzed the 170 compound heterozygous patients (typical SLOS patients) with geographic/ethnicity data from **Table 17** in **Figure 26C, Figure 26D, Figure S1.**

6.4.2 Exon 'Hotspots' Distinguish Among Continents and Between North and South

Figure 26C illustrates how specific DHCR7 mutations occur commonly among certain ethnicities. For example, R352Q and G303R occur almost exclusively among Asian populations (Asian: Korean, Japanese, Japanese-Dutch). Other mutations, e.g., G147D, occur among Southern Europeans (S. Europe: French, Italian, Spanish, Portuguese, Greek). There are some distinctions between Northern and Southern Europeans in terms of DHCR7 mutations. For instance, the null mutation IVS8-1G>C occurs more frequently among Southern Europeans whereas the null mutation W151X occurs more frequently in Northern Europeans. I also found

exon ‘hotspots’ for the pairs of SLOS-inducing DHCR7 mutations. The combination of exons 4 and 9, and also 6 and 9 tended to occur frequently across the whole cohort (see ‘All’ in **Figure 26D).** When I stratified by ethnic group, I found that mutations in exons 4 and 9 or the exon 4-9 ‘hotspot’ were more frequent among Southern Europeans while the exon 6-9 hotspot occurred frequently in Northern Europeans. I observed marked differences for the Asian population with the exon 8-9 hotspot occurring more frequently. Individuals from South America (S. America: Brazil, United States of America-Hispanic) had 3 patterns that were equally frequent, the exon 4-9 hotspot (also common among S. Europeans), the exon 7-9 hotspot (unique to S. Americans), and the exon 8-9 hotspot (common among Asians). Importantly, the frequent N. European hotspot (exon 6-9 hotspot) was absent from the S. American result.

A clear relationship exists between geography/ethnicity and specific DHCR7 mutations (**Figure 26D)** and has been described previously (Al-Owain et al., 2012; Nezarati et al., 2002; Nowaczyk et al., 2004a; Oh et al., 2014). In addition, specific exon ‘hotspots’ also vary by geographic location/ethnicity (**Figure 26D**).

6.5 Pathway Links DHCR7, Vitamin D Synthesis, and Cholesterol Synthesis

6.5.1 Evolutionary Advantage for DHCR7 Mutations

Heterozygote advantage among carriers was proposed previously (Nowaczyk et al., 2006) based on the relationship between DHCR7 and vitamin D. Heterozygote carriers have lower amounts of functional DHCR7 resulting from their carrier state(Fitzky et al., 1998). This would lower the amount of available 7-dehydrocholesterol that is converted to cholesterol. Subsequently, levels of 7-dehydrocholesterol would increase allowing more available 7-dehydrocholesterol to be converted to vitamin D upon exposure to ultraviolet B light.

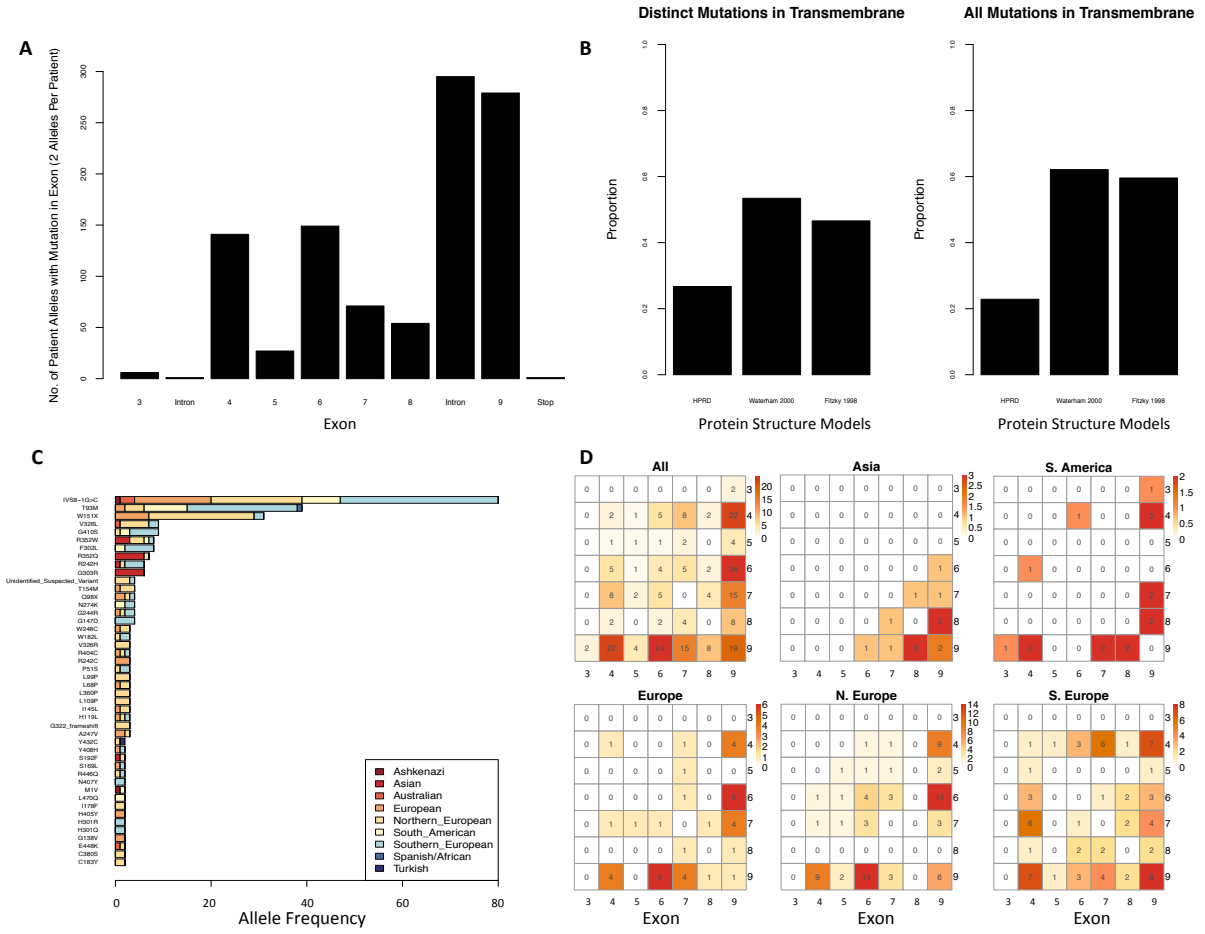


Figure 26. Certain Exons and Functional Regions Are Enriched for SLOS-Inducing Mutations in DHCR7 and Mutation Spectrum Varies By Region and Ethnicity. The overall exon distribution of alleles is shown in **Figure 26A** for all 1024 alleles. Notice that the intronic null mutation (renders exon 9 non-existent) is the most common followed by mutations in exon 9. Exons 4 and 6 also feature prominently in SLOS-inducing mutations. **Figure 26B** depicts the proportion of distinct DHCR7 mutations (left) and overall mutations (right) that occur in the trans-membrane region of the protein. The 3D protein structure for DHCR7 has yet to be published. However, three different models have been described indicating transmembrane domains including the Human Protein Reference Database (HPRD), Waterham et al. (Waterham and Wanders, 2000) and Fitzky et al. (Fitzky et al., 1998) Note that the Waterham model (Waterham and Wanders, 2000) results in the largest proportion (62.109%) of SLOS-inducing mutations being flagged as occurring in the transmembrane domain. **Figure 26C** shows the allele frequency distribution for patients with ethnicity or country of origin information (170 patients had this information available). Some ethnicities only had one patient including, Ashkenazi, Turkish and Spanish/African (a patient with mixed ancestry). **Figure 26D** contains the exon locations for each compound heterozygote pair (each patient has 2 mutations in DHCR7). Note that different ethnicities or countries of origin have different mutation patterns. For example Asian patients frequently have one mutation in exon 8 and one mutation in exon 9; whereas patients from Europe (unspecified lower left-hand corner of **Figure 26D)** or Northern Europe tend to have one mutation in exon 6 and other in exon 9. Patients from Southern Europe frequently had two different mutations in exon 9 or one mutation in exon 4 and one in exon 9. **Ethnicity groupings:** Europe (European But Not Otherwise Specified and N-S Europeans); N. Europe (Dutch, Hungarian, Polish, German, Austrian, United Kingdom, Irish); S. Europe (French, Italian, Spanish, Portuguese, Greek); S. America (Brazil, United States of America-Hispanic); Asia (Korean, Japanese, Japanese-Dutch).

Table 17. Compilation of SLOS DHCR7 Genotypes from 229 Patients Extracted from 21 Studies

Genotype		Prenatal Lethal?	Reference(s)	No. of Patients (N=229)	Incidence (%)
Homozygous					
IVS8-1G>C*	IVS8-1G>C	Yes / or shortly after birth	(Evans et al., 2001; Ginat et al., 2004; Goldenberg et al., 2003; Jira et al., 2001; Lanthaler et al., 2013; Scalco et al., 2005)	18	7.860
W151X	W151X	Yes	(Jezela-Stanek et al., 2010; Lanthaler et al., 2013)	8	3.493
R352Q	R352Q	No	(Al-Owain et al., 2012; Matsumoto et al., 2005; Oh et al., 2014)	6	2.620
T93M	T93M	No	(Cardoso et al., 2005; De Brasi et al., 1999; Nowaczyk et al., 2004a)	4	1.747
R352W	R352W	No	(Oh et al., 2014)	2	0.873
P467L	P467L	No	(Nezarati et al., 2002)	2	0.873
E448K	E448K	No	(Anstey et al., 2005; De Brasi et al., 1999)	2	0.873
IVS8-1G>T	IVS8-1G>T	Yes	(Lanthaler et al., 2013)	1	0.437
G963 ↓134bp Frameshift	G963 ↓134bp Frameshift	Died Shortly After Birth	(Waterham et al., 1998)	1	0.437
L109P	L109P	No	(Jezela-Stanek et al., 2010)	1	0.437
N287K	N287K	No	(Al-Owain et al., 2012)	1	0.437
R446Q	R446Q	NA[†]	(Goldenberg et al., 2003)	1	0.437
R352L	R352L	No	(Al-Owain et al., 2012)	1	0.437
Total Homozygous				**48**	**20.961**
Compound Heterozygous					
IVS8-1G>C	T93M	No	(Anstey et al., 2005; Cardoso et al., 2005; Ginat et al., 2004; Nowaczyk et al., 2004a; Patrono et al., 2002; Scalco et al., 2005)	21	9.170
IVS8-1G>C	W151X	No	(Lanthaler et al., 2013)	7	3.057
IVS8-1G>C	G410S	No	(Goldenberg et al., 2003)	6	2.620
IVS8-1G>C	T154M	Died Shortly After Birth/No	(Ginat et al., 2004; Jira et al., 2001)	3	1.310
G147D	F302L	Yes for 1, NA for 2	(Goldenberg et al., 2003)	3	1.310
R352Q	G303R	No/Died Shortly After Birth	(Matsumoto et al., 2005; Oh et al., 2014)	3	1.310
IVS8-1G>C	C183Y	Died Shortly After Birth in Some Cases	(Jira et al., 2001)	2	0.873
T93M	Q98X	No	(Cardoso et al., 2005; Witsch-Baumgartner et al., 2005)	2	0.873
W151X	V326L	No	(Fitzky et al., 1998)	2	0.873
W151X	C380S	No	(Fitzky et al., 1998)	2	0.873
V326R	L68P	No	(Witsch-Baumgartner et al., 2005)	2	0.873
IVS8-1G>C	M1V	No	(Scalco et al., 2005; Witsch-Baumgartner et al., 2005)	2	0.873
IVS8-1G>C	P51S	No	(Fitzky et al., 1998; Goldenberg et al., 2003)	2	0.873
IVS8-1G>C	L99P	No	(Anstey et al., 2005; Fitzky et al., 1998)	2	0.873
IVS8-1G>C	S169L	Died Shortly After Birth/No	(Ginat et al., 2004; Goldenberg et al., 2003)	2	0.873
IVS8-1G>C	A247V	No	(Ginat et al., 2004)	2	0.873

IVS8-1G>C	F302L	Possibly[††]	(Goldenberg et al., 2003)	2	0.873
IVS8-1G>C	V326L	No	(Evans et al., 2001; Goldenberg et al., 2003)	2	0.873
W151X	I145L	No	(Jezela-Stanek et al., 2010)	2	0.873
W151X	L157P	No	(Fitzky et al., 1998; Jezela-Stanek et al., 2010)	2	0.873
W151X	V326L	No	(Jezela-Stanek et al., 2010)	2	0.873
H119L	G244R	Died Shortly After Birth in Some Cases	(Jira et al., 2001; Waterham et al., 1998)	2	0.873
H301Q	R242H	No	(Goldenberg et al., 2003)	2	0.873
F302L	L470Q	No	(Ginat et al., 2004)	2	0.873
G303R	R352W	No	(Oh et al., 2014)	2	0.873
H405Y	G138V	No	(Waye et al., 2005)	2	0.873
T93M	W151X	No	(De Brasi et al., 1999; Patrono et al., 2002)	2	0.873
W151X	R352W	No	(Jezela-Stanek et al., 2010)	2	0.873
T93M	G147D	No	(Goldenberg et al., 2003)	1	0.437
T93M	T154R	No	(Scalco et al., 2005)	1	0.437
T93M	S192F	No	(Witsch-Baumgartner et al., 2005)	1	0.437
T93M	R228W	No	(Witsch-Baumgartner et al., 2005)	1	0.437
T93M	D234Y	No	(Nowaczyk et al., 2004a)	1	0.437
T93M	R242H	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
T93M	G244R	No	(Patrono et al., 2002)	1	0.437
T93M	N274K	No	(Cardoso et al., 2005)	1	0.437
T93M	V281M	No	(Nowaczyk et al., 2004a)	1	0.437
T93M	F302L	No	(Nowaczyk et al., 2004a)	1	0.437
T93M	G410S	No	(Scalco et al., 2005)	1	0.437
T93M	720-735del Frameshift	No	(Fitzky et al., 1998)	1	0.437
T93M	IVS5+4 del	No	(De Brasi et al., 1999)	1	0.437
IVS8-1G>C	A50N	No	(Witsch-Baumgartner et al., 2005)	1	0.437
IVS8-1G>C	P51H	No	(Anstey et al., 2005)	1	0.437
IVS8-1G>C	Q98X	No	(Lanthaler et al., 2013)	1	0.437
IVS8-1G>C	L109P	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
IVS8-1G>C	H119L	No	(Goldenberg et al., 2003)	1	0.437
IVS8-1G>C	H119fsX8	No	(Witsch-Baumgartner et al., 2005)	1	0.437
IVS8-1G>C	F174S	No	(Cardoso et al., 2005)	1	0.437
IVS8-1G>C	W182C	No	(Goldenberg et al., 2003)	1	0.437
IVS8-1G>C	W182L	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
IVS8-1G>C	K198E	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
IVS8-1G>C	F235S	No	(Waye et al., 2005)	1	0.437
IVS8-1G>C	R242H	NA[†]	(Goldenberg et al., 2003)	1	0.437
IVS8-1G>C	W248C	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
IVS8-1G>C	S254A	No	(Goldenberg et al., 2003)	1	0.437
IVS8-1G>T	F255L	Died Shortly After Birth	(Jira et al., 2001)	1	0.437
IVS8-1G>C	I297T	No	(Waye et al., 2005)	1	0.437
IVS8-1G>C	H301R	No	(Cardoso et al., 2005)	1	0.437
IVS8-1G>C	P329L	No	(Patrono et al., 2002)	1	0.437
IVS8-1G>C	356delH	No	(Evans et al., 2001)	1	0.437
IVS8-1G>C	G366V	No	(Szabó et al., 2010)	1	0.437
IVS8-1G>C	C380Y	No	(Ginat et al., 2004)	1	0.437
IVS8-1G>C	R404C	Died Shortly After Birth	(Goldenberg et al., 2003)	1	0.437
IVS8-1G>C	Y432C	No	(Witsch-Baumgartner et al., 2005)	1	0.437
IVS8-1G>C	R446Q	No	(Patrono et al., 2002)	1	0.437

IVS8-1G>C	E448K	No	(Evans et al., 2001)	1	0.437
IVS8-1G>C	R450L	No	(Anstey et al., 2005)	1	0.437
IVS8-1G>C	F475S	No	(Witsch-Baumgartner et al., 2005)	1	0.437
IVS8-1G>C	Unidentified Suspected Variant	No	(Cardoso et al., 2005)	1	0.437
W151X	L109P	Died Shortly After Birth	(Jezela-Stanek et al., 2010)	1	0.437
W151X	N146K	No	(Jezela-Stanek et al., 2010)	1	0.437
W151X	I178F	No	(Witsch-Baumgartner et al., 2005)	1	0.437
W151X	W248C	No	(Jezela-Stanek et al., 2010)	1	0.437
W151X	G347S	No	(Witsch-Baumgartner et al., 2005)	1	0.437
					0.437
W151X	R352Q	No	(Jezela-Stanek et al., 2010)	1	0.437
W151X	L360P	No	(Witsch-Baumgartner et al., 2005)	1	0.437
W151X	R446Q	Died Shortly After Birth	(Jezela-Stanek et al., 2010)	1	0.437
W151X	Unidentified Suspected Variant	Yes	(Jezela-Stanek et al., 2010)	1	0.437
W151X	G322 frameshift	No	(Jezela-Stanek et al., 2010)	1	0.437
P51S	N274K	Died Shortly After Birth	(Goldenberg et al., 2003)	1	0.437
G322 frameshift	Unidentified Suspected Variant	No	(Jezela-Stanek et al., 2010)	1	0.437
V326R	L360P	No	(Witsch-Baumgartner et al., 2005)	1	0.437
Q98X	Unidentified Suspected Variant	No	(Witsch-Baumgartner et al., 2005)	1	0.437
L99P	G410S	No	(Fitzky et al., 1998)	1	0.437
T154M	Y219D	No	(Jezela-Stanek et al., 2010)	1	0.437
I178F	R242H	No	(Witsch-Baumgartner et al., 2005)	1	0.437
W182L	E224K	No	(Witsch-Baumgartner et al., 2005)	1	0.437
P243R	Y408H	Died Shortly After Birth	(Goldenberg et al., 2003)	1	0.437
A247V	R404C	No	(Fitzky et al., 1998)	1	0.437
V273G	Y432C	No	(Witsch-Baumgartner et al., 2005)	1	0.437
					0.437
H301R	W182L	No	(Cardoso et al., 2005)	1	0.437
					0.437
V326L	G244R	No	(Goldenberg et al., 2003)	1	0.437
V326L	R352W	No	(Fitzky et al., 1998)	1	0.437
V326L	E448K	No	(Jezela-Stanek et al., 2010)	1	0.437
Intron: G963 insertion of 134bp Frameshift	W248C	No	(Waterham et al., 1998)	1	0.437
T93M	N407Y	No	(De Brasi et al., 1999)	1	0.437
R352W	N407Y	Died Shortly After Birth	(Goldenberg et al., 2003)	1	0.437
R352W	K376R fs*37	No	(Oh et al., 2014)	1	0.437
R352W	L317R	No	(Scalco et al., 2005)	1	0.437
G322 Frameshift	L109P	Yes	(Jezela-Stanek et al., 2010)	1	0.437
V330M	R363C	No	(Patrono et al., 2002)	1	0.437
R352Q	R242H	No	(Matsumoto et al., 2005)	1	0.437
R352Q	X476Q	No	(Matsumoto et al., 2005)	1	0.437
R352Q	S192F	No	(Matsumoto et al., 2005)	1	0.437
P227S	G303R	No	(Oh et al., 2014)	1	0.437
L68P	R404C	No	(Waye et al., 2005)	1	0.437
Q107H	C444Y	No	(Ginat et al., 2004)	1	0.437
I145L	Y408H	No	(Waye et al., 2005)	1	0.437
N274K	V466A	No	(Scalco et al., 2005)	1	0.437
N274K	G410S	No	(Scalco et al., 2005)	1	0.437
R242C	G344R	No	(Waye et al., 2005)	1	0.437

R242C	W177R	No	(Ginat et al., 2004)	1	0.437
R242C	H426P	No	(Waye et al., 2005)	1	0.437
E288K	I215N	No	(Romano et al., 2005)	1	0.437
682-683 insert-C	98-184 deletion	No	(Ginat et al., 2004)	1	0.437
Total Compound Heterozygous				**172**	**75.109**
True Heterozygous					
R352W	None	No	(De Brasi et al., 1999; Fitzky et al., 1998)	3	1.310
T93M	None	No	(De Brasi et al., 1999)	2	0.873
V326L	None	No	(Fitzky et al., 1998)	2	0.873
S113C	None	No	(Waye et al., 2005)	1	0.437
E448K	None	No	(Patrono et al., 2002)	1	0.437
Total True Heterozygous				**9**	**3.930**

*Annotated as: c.964-1G>C in some literature articles

† Not Applicable, medical termination of pregnancy occurred (Goldenberg et al., 2003)

†† In one case natural miscarriage occurred, in second case medical termination of pregnancy occurred (Goldenberg et al., 2003)

Died Shortly After Birth= Less than 1 year old

This would protect individuals from developing rickets (a vitamin D deficiency condition that results in bone softening and malformations) and osteomalacia (an adult form of rickets) (Nowaczyk et al., 2006).

Certain DHCR7 variants were found to be evolutionarily favored in Northern climates, including Europe and Asia (Kuan et al., 2013) suggesting a selective pressure and a plausible heterozygote advantage for those variants. Researchers found certain DHCR7 variants were associated with ***lower*** vitamin D levels in the general population (Wang et al., 2010) and among individuals with polycystic ovary syndrome (Wehr et al., 2011). This further links vitamin D levels with DHCR7 variants.

Historically, there would have been an evolutionary pressure to maximize the small amount of sunlight available for vitamin D production in Northern climates due to lower sunlight exposure. Additionally, benefits from prenatal vitamin D supplementation are also modulated by geographic location and other factors (Karras et al., 2015). Therefore biological mechanisms that maximize vitamin D production and absorption would be selected for in the North. A heterozygote advantage for individuals with one DHCR7 mutation would therefore exist to allow most of the body's 7-dehydrocholesterol to be converted into vitamin D (and not cholesterol). This has been used to explain why Northern populations have a higher prevalence of SLOS (Kelley and Hennekam, 2000).

Carriers also are hypothesized to have a reproductive advantage because of the reduced fetal death due to rachitic cephalopelvic disproportion (Kelley and Hennekam, 2000; Opitz et al., 2002). This is again related to increased vitamin D production causing improved bone formation. Improved hip formation is thought to enable females to produce more offspring and thereby provide another carrier advantage (Kelley and Hennekam, 2000; Opitz et al., 2002).

Additionally, vitamin D has been shown to increase chances of pregnancy for infertile women receiving in vitro fertilization (Ozkan et al., 2010). Vitamin D has many pleiotropic roles in reproductive outcomes and can affect both male and female fertility (Lerchbaum and Obermayer-Pietsch, 2012). Therefore, there could be multiple mechanisms that could explain why evolution favors vitamin D production in Northern climates through mutations in DHCR7 (and other genes).

6.5.2 Biological Mechanism: DHCR7's Effect on Vitamin D and Cholesterol Synthesis

The biological mechanism that connects DHCR7 mutations (SLOS carriers) and Northern climate involves regulation of 7-dehydrocholesterol, vitamin D, and cholesterol. Typically, 7-dehydrocholesterol can be converted into either cholesterol or vitamin D (cholecalciferol). The conversion of 7-dehydrocholesterol to vitamin D occurs in the skin upon exposure to ultraviolet B light (290-320 nm) (Deeb et al., 2007). Under normal sunlight conditions, only around 15% of available 7-dehydrocholesterol will be converted to vitamin D while excess 7-dehydrocholesterol in the skin will be converted to inert compounds for degradation (Norman, 1998; Webb et al., 1988). DHCR7 is the sole enzyme used to convert 7-dehydrocholesterol to cholesterol (Moebius et al., 1998). A shortage of DHCR7 would decrease the body's ability to convert 7-dehydrocholesterol into cholesterol thereby causing a buildup of 7-dehydrocholesterol, which could be converted into vitamin D in the skin. Each piece of this mechanistic pathway is depicted in **Figure 27** along with the corresponding literature references.

An indirect negative feedback loop, established in laboratory studies, allows vitamin D levels to regulate DHCR7 activity and prevent hypervitaminosis D (i.e., toxically high vitamin D levels) (Zou and Porter). SLOS patients have high 7-dehydrocholesterol levels. Therefore, from **Figure 27** and the literature, one would expect them to have high vitamin D levels. However, vitamin D

levels in SLOS patients are reported to be low (Rossi et al., 2005). This could be due to many lifestyle factors that occur when a patient is seriously ill with SLOS. One possible biological explanation that fits with the literature-derived mechanism (**Figure 27)** is that SLOS patients spent less time outdoors thereby reducing their ultraviolet B exposure (another critical requirement for vitamin D synthesis) (Rossi et al., 2005).

6.6 Genetic Understanding of SLOS-Inducing DHCR7 Mutations: Implications for Future Work

6.6.1 Changing The Understanding of SLOS Genetics Using Large-Scale Genomics Studies

Using ExAC (Lek et al., 2015), I was able to compare the frequencies of DHCR7 SLOS-inducing mutations in the SLOS population (extracted from the literature) and the assumed-healthy ExAC population. I found that there were many differences. To illustrate some of these differences, I plotted the overall incidence of DHCR7 mutations in SLOS against the incidence in the healthy population (ordered by healthy population). Several DHCR7 mutations are much more frequent among SLOS individuals then expected given the background population rate (**Figure 28).** For example, T93M is the second most frequent DHCR7 mutation in SLOS patients, but it is comparatively rare among the ExAC population (**Figure 28**) (Lek et al., 2015). This could be due to sampling bias differences between the SLOS cohort and the ExAC population. For example, T93M is thought to be the founder SLOS mutation and is common in Italy, Spain and Portugal (Cardoso et al., 2005; Kozák et al., 2000; Nowaczyk et al., 2004a; Witsch-Baumgartner et al., 2001; Witsch-Baumgartner et al., 2005). The ExAC population may be under-represented for Southern Europeans. In ExAC they distinguish a 'Finnish' European cohort from a non-Finnish European cohort (representing Northern American peoples and other Europeans), but it is unclear how many Southern Europeans are represented in that cohort.

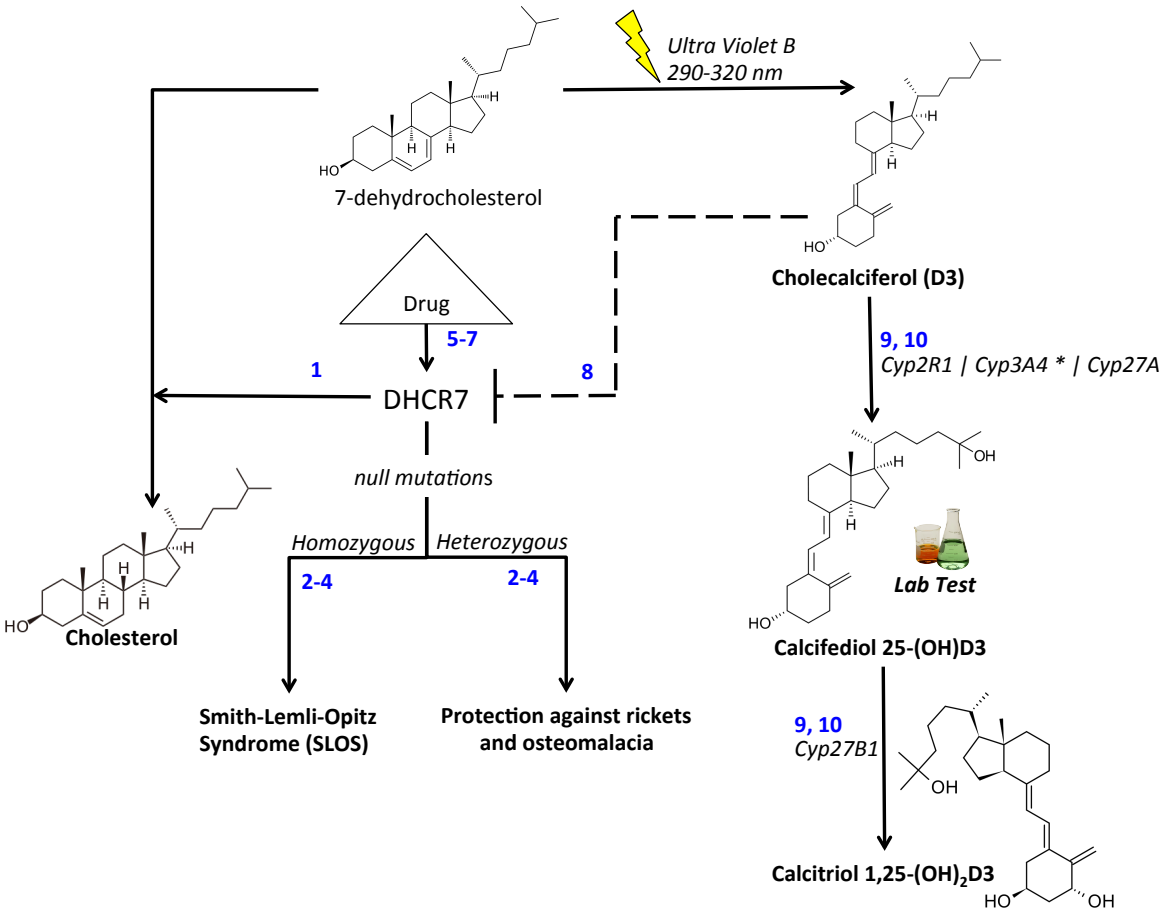


Figure 27. Literature-Derived Pathway Illustrates How 7-DeHydroCholesterol Reductase Effects Vitamin D Production By Removing 7-Dehydrocholesterol and the Effects of Drugs on this Pathway. Drugs that enhance DHCR7 (7-DehydroCholesterol Reductase) function result in reduction in vitamin D production. This occurs because DHCR7 causes more 7-dehydrocholesterol to be converted to cholesterol. This indirectly inhibits vitamin D production by reducing the amount of 7-dehydrocholesterol that can be converted into cholecalciferol in the skin. Homozygous null mutations in DHCR7 completely inhibit functionality and result in patients with Smith-Lemli-Opitz Syndrome (SLOS) with heterozygous patients showing increased protection against rickets and osteomalacia. The dashed line indicates a feedback inhibition loop.

Reference Legend:

1 (Moebius et al., 1998); **2** (Wang et al., 2010); **3** (Ahn et al., 2010); **4** (Tint et al., 1994); **5** (Lauth et al., 2010); **6** (Fernø et al., 2005); **7** (Raeder et al., 2006); **8** (Zou and Porter); **9** (Kuan et al., 2013); **10** (Deeb et al., 2007)

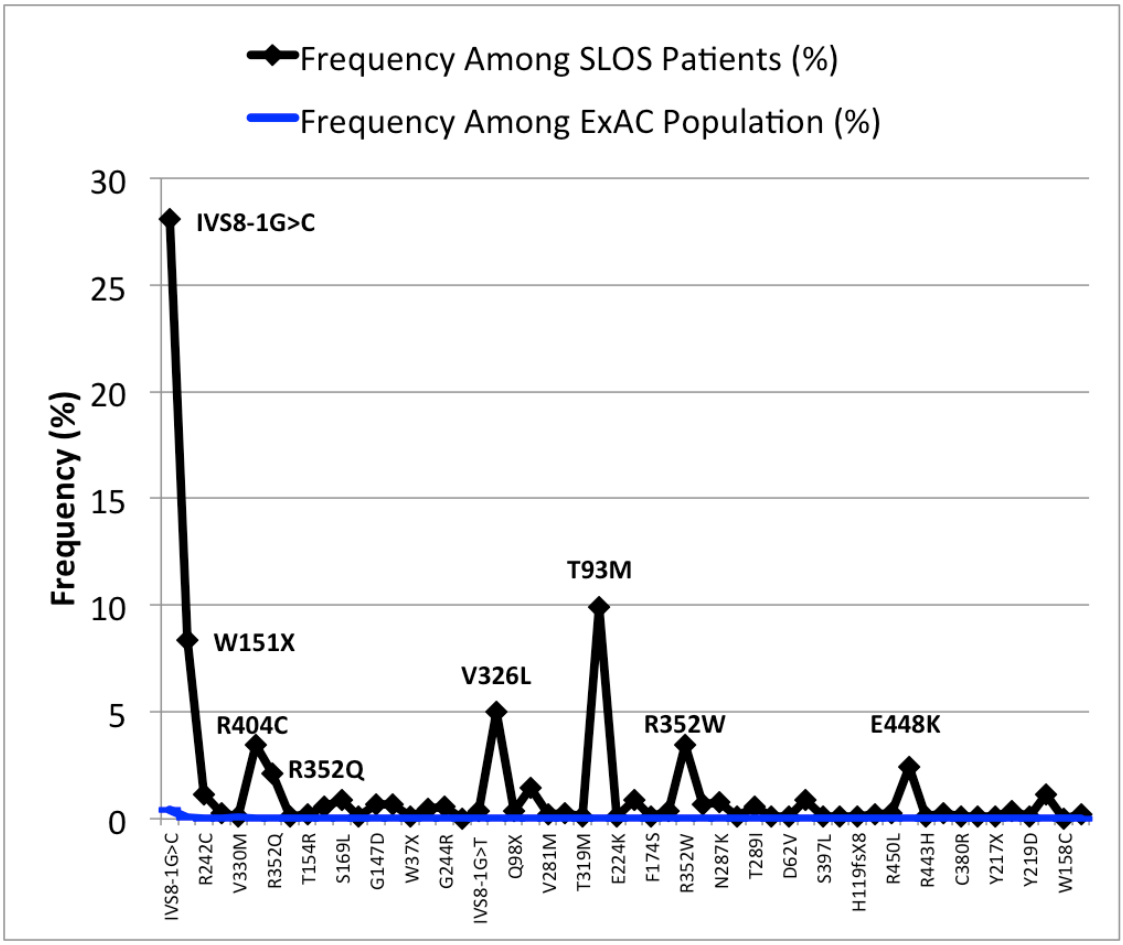


Figure 28. DHCR7 SLOS-Inducing Mutation Frequency in SLOS Patients vs. Frequency from ExAC Population. Some mutations occurred more frequently in the SLOS population given their frequency in the healthy population (e.g., T93M, V326L, R404C, R352W). A possible explanation for this could be that these mutations occur in certain ethnic populations and those populations are under-represented in the ExAC population.

Table 18. DHCR7 Mutations Predicted to Be Damaging from ExAC Cohort (60,706 Individuals) Includes Both Known SLOS-Inducing Mutations and Unknown Mutations

Implicated in SLOS?	Position	RSID	Protein	Transcript	Ann.	Overall Allele Count (Freq. %)
Yes	71146886	rs138659167		c.964-1G>C	splice acceptor	386 (4.2 X 10^{-1})
Yes	71152447	rs11555217	p.Trp151Ter	c.452G>A	stop gained	82 (6.8 X 10^{-2})
No	71146229	rs115338563	p.Arg207Ter	c.619C>T	stop gained	6 (5.4 X 10^{-2})
No	71146487	rs147850435	-	c.611+1G>A	splice donor	8 (6.8 X 10^{-3})
Yes	71146886	rs138659167	-	c.964-1G>T	splice acceptor	6 (6.5 X 10^{-3})
Yes	71155249	.	p.Trp37Ter	c.111G>A	stop gained	3 (5.7 X 10^{-3})
Yes	71155068	rs104886039	p.Gln98Ter	c.292C>T	stop gained	5 (4.8 X 10^{-3})
AA*	71152354	.	p.Trp182Ter	c.545G>A	stop gained	5 (4.1 X 10^{-3})
Yes	71146904	.	p.Pro149Hisfs Ter163	c.442_445dupACCC	frameshift	2 (2.7 X 10^{-3})
No	71152488	.	-	c.413-2A>G	splice acceptor	2 (2.3 X 10^{-3})
AA	71146937	.	p.Leu138Cysfs Ter10	c.412delC	frameshift	1 (2.0 X 10^{-3})
No	71146790	.	p.Trp187Ter	c.560G>A	stop gained	2 (1.7 X 10^{-3})
AA	71146782	.	p.His356Thrfs Ter57	c.1066delC	frameshift	2 (1.7 X 10^{-3})
No	71155195	.	p.Tyr55Ter	c.165C>G	stop gained	1 (1.1 X 10^{-3})
No	71153313	.	p.Val134Cysfs Ter90	c.400_408delGTGACTCCTinsT	frameshift	1 (9.3 X 10^{-4})
AA	71153346	.	p.Lys120Thrfs Ter2	c.359_375delAGTTTCTACCCGGCTACinsC	frameshift	1 (8.7 X 10^{-4})
Yes	71153365	.	p.His119Ilefs Ter8	c.355_356delCAinsA	frameshift	1 (8.6 X 10^{-4})
AA	71146791	.	p.Val353Trpfs Ter60	c.1057delG	frameshift	1 (8.5 X 10^{-4})
No	71146559	rs140791666	p.Tyr430Ter	c.1290C>G	stop gained	1 (8.5 X 10^{-4})
AA	71146709	.	p.Cys380Ter	c.1140C>A	stop gained	1 (8.4 X 10^{-4})
Yes	71150105	.	p.Tyr217Ter	c.651C>A	stop gained	1 (8.3 X 10^{-4})
No	71148951	.	p.Trp290Ter	c.870G>A	stop gained	1 (8.3 X 10^{-4})
No	71155917	.	p.Gln28Ter	c.82C>T	stop gained	1 (8.3 X 10^{-4})
No	71155938	.	p.Asp21Argfs Ter42	c.60_61insA	frameshift	1 (8.3 X 10^{-4})

* AA: Indicates that another mutation at that same amino acid (AA) position number has been implicated in SLOS. However, the mutation is different from the one found by ExAC.

Therefore, some of these differences are likely due to sampling bias.

Additionally, I was able to identify DHCR7 mutations that were predicted to be damaging using ExAC. This revealed 24 mutations (**Table 18).** Eight of those mutations were already known to be implicated in SLOS and were contained in the compendium. Six more mutations occurred at the same amino acid position of another mutation that has been implicated in SLOS (demonstrating the importance of that amino acid position in SLOS). However, 12 mutations were never reported as being found in SLOS patients. There are two potential reasons for this: 1) the mutation is not disease causing (this is difficult to ascertain without the protein structure); and 2) the mutation only exists in an under-studied research population. For example, one mutation (R207X) was found to have 2.2% frequency among the healthy African population from ExAC. This means that R207X is polymorphic among Africans. However, this mutation has never been implicated in SLOS. Its possible that Africans are not often diagnosed with SLOS and then sequenced because they often come from resource-poor environments. The 12 mutations in **Table 18** that have never been implicated in SLOS are worthy of further investigation by researchers to ascertain whether they are deleterious. If these mutations were deleterious, the next research question would be why African populations are under-diagnosed for SLOS.

6.6.2 Implications for Causality

Hill established a set of nine criteria for determining whether an association was causal or not (Hill, 1965). I will focus discussion on three of these criteria with regards to DHCR7: strength, plausibility and coherence. Strength is difficult to assess in SLOS because the DHCR7 mutations are often only found among SLOS patients (and researchers are biased to look specifically for mutations in DHCR7). As can be seen from **Table 18** the incidence of these mutations in the

ExAC population is very low. This helps to strengthen the belief in the relationship between DHCR7 mutations and SLOS, especially for mutations that are disproportionately high among SLOS patients (e.g., T93M in **Figure 28)**. However, the fact that some mutations are very high among SLOS patients and very low among the ExAC population could be due to sampling bias between the two populations. Some DHCR7 mutations were found in the ExAC population that is predicted to be functionally deleterious. However, not all of these have been reported to be SLOS-Inducing. This could be due to ethnic differences between the two datasets (i.e., certain groups are under-studied and therefore some of the predicted damaging mutations are actually un-reported SLOS-Inducing mutations) or it could be because some predicted damaging mutations are not functionally damaging (this would be easier to determine if the structure of DHCR7 were known). Because the protein structure of DHCR7 remains unknown it is difficult to predict for certain whether a mutation is deleterious or not.

A paper by Lanthaler et al. (Lanthaler et al., 2013) found that certain maternal genetic signatures could help rescue the SLOS phenotype by increasing the amount of cholesterol passed to the offspring via the placenta. This variation in the maternal genome could perhaps explain some of the discrepancies between the literature-derived SLOS compendium and the ExAC population. For example, in addition to the requirement that two DHCR7 mutations be inherited, it may also be necessary to have a certain placenta state (or placental gene mutation) to acquire the disease. This could perhaps explain why some common DHCR7 mutations from ExAC are not found to be SLOS-inducing (even though they are predicted to be damaging). However, it is too early to state yet because of sampling bias issues and under-reporting among certain ethnicities. Further investigation could help provide additional coherence to reported DHCR7 mutations in the literature. Therefore, this remains an open area of research.

6.7 Pharmacological Effects of DHCR7 Modulators

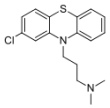
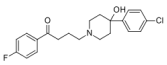
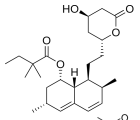
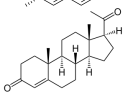
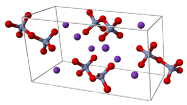
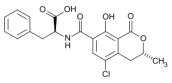
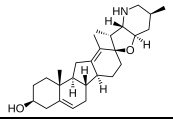
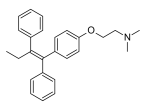
Importantly, pharmaceutical drugs have been shown to modulate DHCR7 activity in various ways. Several anti-psychotic drugs were found to enhance DHCR7 activity (Fernø et al., 2005; Lauth et al., 2010; Raeder et al., 2006). Other drugs, not otherwise known to modulate DHCR7, can cause high 7-dehydrocholesterol in the absence of SLOS (Hall et al., 2013). Currently, pharmaceuticals are being designed to inhibit DHCR7 as a suggested treatment for hepatitis C (Rodgers et al., 2012). However, diverse therapeutic uses exist for targeting DHCR7. In this section, I review all known pharmaceuticals, compounds, chemicals, and toxins known to modulate DHCR7.

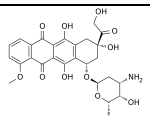
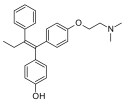
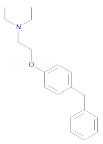
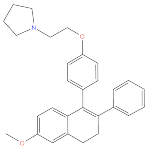
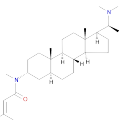
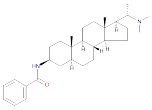
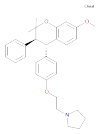
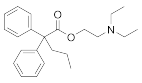
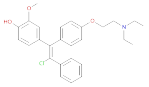
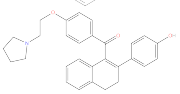
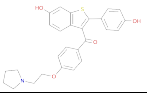
6.7.1 DHCR7 Modulators

6.7.1.1 Expression

I used the Comparative Toxicogenomics Database (CTD) (CTD, 2015; Davis et al., 2015) to retrieve articles describing compounds that modulate DHCR7 expression (data retrieved February 2015). I then manually reviewed the resulting literature references and only retained compounds with direct evidence of modulating DHCR7 expression in studies using human tissue. **Table 19** contains the list of modulators with corresponding references**.** Several other compounds were found in CTD to modulate DHCR7, but failed the requirements listed above these are presented as potential DHCR7 modulators in **Table S3** (of the published paper)**.** Overall I found 5 pharmaceuticals and 2 inorganic compounds increased DHCR7 expression and 3 toxins decreased DHCR7 expression. Antipsychotic drugs, such as haloperidol, clozapine, and chlorpromazine, are known to alter expression in cardiovascular genes (Foley and Mackinnon, 2014) in addition to DHCR7 (increased DHCR7 activity would increase the rate of cholesterol synthesis see **Figure 27**).

Table 19. Chemicals Known to Modulate DHCR7 with Literature References

Chemical Type†	Chemical /Drug	Fetal Risk Category†‡	Reference(s)	Source	Structure
Increases Expression (up-regulates)					
Pharmaceutical Drugs					
Antipsychotic [Dibenzazepine]	Clozapine	B	(Fernø et al., 2005; Ferno et al., 2006; Lauth et al., 2010; Raeder et al., 2006)	Literature Review/ CTD	
Antipsychotic [Phenothiazine]	Chlorpromazine	C	(Fernø et al., 2005; Lauth et al., 2010; Raeder et al., 2006)	Literature Review	
Antipsychotic [Butyrophenone]	Haloperidol	C	(Fernø et al., 2005; Lauth et al., 2010; Raeder et al., 2006)	Literature Review	
Statin [Naphthalene]	Simvastatin	X	(Correa-Cerro et al., 2006; Wassif et al., 2005)	Literature Review/ CTD	
Corpus Luteum Hormone	Progesterone	B	(Wilcox et al., 2007)	Literature Review/ CTD	
Inorganic Compounds					
Inorganic Element [Fullerene]	Nanotubes, Carbon	-	(Park et al., 2014)	Literature Review/ CTD	NA
Hormone Antagonist [Inorganic]	Potassium Dichromate	-	(Guo et al., 2013)	Literature Review/ CTD	
Decreases Expression (down-regulates)					
Toxins					
Industrial byproduct of water treatment	*N*-nitrosodimethylamine (DMN)	Toxin	(Kawata et al., 2007)	Literature Review/ CTD	
Mycotoxin	Ochratoxin A	Toxin	(Hundhausen et al., 2008)	Literature Review/ CTD	
Metabolite of Inorganic Arsenic	Monomethylarsonous acid	Toxin	(Guo et al., 2014)	Literature Review/ CTD	NA
Decreases Protein Activity					
Poisonous steroidal jerveratrum alkaloid	Cyclopamine (11-deoxojervine)	Poison	(Zou and Porter)	Literature Review	
Direct Inhibitors					
Approved Chemotherapeutics					
Chemotherapeutic, antagonist of the estrogen receptor [Benzylidene]	Tamoxifen (IC50=12nM; Kd=1nM)	D	(Bignon et al., 1989; Jordan, 2003; Ruenitz et al., 1989; Teo et al., 1992)	Chembl	

Chemotherapeutic [Daunorubicin]	Doxorubicin (IC50=150-10000nM)	D	(Burke et al., 2004)	Chembl	
Investigational Chemotherapeutics					
Clinical Trials					
Selective Estrogen Receptor Modulator (SERM) (active metabolite of Tamoxifen)	Afimoxifene	Clinical Trials	(Bignon et al., 1989)	Chembl	
Antagonist of intracellular histamine	Tesmilifene	Clinical Trials	(Teo et al., 1992)	Chembl	
Non-steroidal anti-estrogenic [Pyrrolidine]	Nafoxidine	Clinical Trials (in 1978)	(Teo et al., 1992)	Chembl	
Pre-Clinical					
Tamoxifen-induced antiestrogen activity	(+)-Pachysamine B (IC50=600nM)	-	(Chang et al., 1998)	Chembl	
Tamoxifen-induced antiestrogen activity	Epipachysamine D (IC50=20000nM)	-	(Chang et al., 1998)	Chembl	
Antifertility agent, Ormeloxifene is a SERM	Rel-Ormeloxifene	-	(Teo et al., 1992)	Chembl	
Inhibitor of Cytochrome P450 [Fatty Acid, Valerate]	Proadifen	-	(Teo et al., 1992)	Chembl	
Metabolite of Clomiphene [Benzylidene, Stilbene]	3'-Methoxy-4'Hydroxy Clomiphene (IC50=22nM)	-	(Ruenitz et al., 1989)	Chembl	
Selective Estrogen Receptor Modulator (SERM)	Trioxifene	-	(Sharma et al., 1990)	Chembl	
Anti-estrogen [Raloxifene Analog]	LY-117018	-	(Sharma et al., 1990; Teo et al., 1992)	Chembl	

† Chemical Type Determined Using the Anatomical Therapeutic Chemical Classification Browser: http://mor.nlm.nih.gov/RxClass/

†† Fetal Risk Categories are based on Food Drug Administration's Criteria for the United States of America

NA: Not Available

Note: Structure diagrams are from Chembl and/or Wikipedia

6.7.1.2 Inhibition

I used ChEMBL, a freely-available semi-curated database of bioactive molecules (Bento et al., 2014), to find all pharmacological compounds that directly inhibit DHCR7 (**Table 19).** This includes approved chemotherapeutics, and investigational chemotherapeutics either in clinical or preclinical trials. When available I report IC-50 values (**Table 19**). Drugs with IC-50 values less than 10,000nM or 10 uM are considered potentially pharmacologically relevant inhibitors (Hopkins and Groom, 2002; Keller et al., 2006). I included tamoxifen (IC50=12nM) and doxorubicin (IC50=150nM) as FDA-approved DHCR7 inhibitors.

DHCR7 inhibitors are used to treat cancer, typically breast cancer. For example, nafoxidine inhibits DHCR7 (Teo et al., 1992) and is a known oestrogen antagonist. Several clinical trials were performed in the 1970s for nafoxidine (Engelsman et al., 1975; Group, 1972; Jain et al., 1977; Steinbaum et al., 1978). The major reaction to the drug consisted of dermatitis (55% of patients) exacerbated by sunlight (Steinbaum et al., 1978). This indicates that the drug effected patients' dermal response to sunlight exposure (which could be related to its inhibition of DHCR7 and dysregulation of vitamin D synthesis pathways).

6.7.1.3 Summary Modulators

Overall I found that five approved pharmaceuticals increase DHCR7 expression while 2 inhibit DHCR7. Two inorganic compounds increase DHCR7 expression, while 3 toxins (including a metabolite of arsenic) decrease DHCR7 expression and one toxin decreases protein activity. Ten pharmaceutical inhibitors of DHCR7 are in various stages of clinical development. This includes 7 inhibitors in the pre-clinical stage and 3 inhibitors in the clinical trials stage (**Table S4,** published paper). DHCR7 Inhibitors are used or will be used (if approved) for cancer treatment.

6.7.2 Investigating Fetal Outcomes Following Prenatal Exposure to DHCR7 Modulators

6.7.2.1 Overview

The biological mechanisms underlying pharmacological teratogenicity and adverse fetal outcomes are complex involving many genetic and environmental factors. In the United States of America, major developmental defects occur in approximately 3-5% of live-born children (Finnell, 1999). Among these defects between 2-3% are classifiable as teratogen-induced meaning that the defect occurred due to a known or suspected prenatal environmental exposure (Finnell, 1999). It is further estimated that <1% of teratogenic effects have a pharmacological origin (Beckman and Brent, 1984).

Interactions between pharmaceutical drugs and genetics could result in increased teratogenicity among certain individuals. For example, holoprosencephaly clusters in families (Roach et al., 1974), can result from mutations in DHCR7 (Shim et al., 2004), and can also occur after exposure to certain drugs, namely lovastatin (Edison and Muenke, 2004a; 2005). It is reasonable to assume that various combinations of those factors (e.g., drug exposure, DHCR7 mutations, familial risk factors) could result in increased likelihood for teratogenic effects. However, for the purposes of this review, I focus on fetal outcomes resulting from prenatal drug exposure to DHCR7 modulators.

Because of the importance of functioning DHCR7 during fetal development, I decided to investigate the outcomes of prenatal exposure to two approved pharmaceutical DHCR7 inhibitors (**Table S4**, published paper) and compare them against five approved pharmaceutical drugs that increase DHCR7 expression. I was interested in the fetal outcomes of both DHCR7 inhibitors and those that increase expression to determine if DHCR7 inhibitors resulted in increased detrimental fetal outcomes. I made this hypothesis because DHCR7 inhibition results

in a pharmaceutically induced SLOS-like fetal development environment and therefore prenatal exposure to these inhibitors should result in SLOS like adverse outcomes.

6.7.2.2 PubMed Literature Review on Fetal Toxicity

Using PubMed, I retrieved all relevant observational studies on fetal outcomes due to prenatal exposure of each drug. The full details regarding the semi-automated query and retrieval process for included articles are described in the **Supplemental** of the published paper. I reviewed each observational study carefully recording the number of pregnancy outcomes per drug treatment. Seven main pregnancy outcomes were used: born healthy, spontaneous abortion (or miscarriage), elective termination (or induced abortion), neonatal death (died within 1st week of birth), still birth (or intrauterine death), ectopic pregnancy, and major/minor fetal malformations or defects or congenital anomalies. Whenever possible I included information regarding the trimester of drug exposure, as this is critical information for understanding teratogenic effects, realizing that SLOS-like symptoms would result mainly from first-trimester exposure (when cholesterol production is critical for proper structural formation). As controls, I included one known teratogen and one known pregnancy-safe drug. I selected isotretinoin (trade name: Accutane) as the known teratogen and levothyroxine as the known pregnancy-safe drug (FDA category A).

My method uses data gleaned from literature reports on fetal outcomes following prenatal drug exposure. Therefore, this data suffers from reporting bias. This occurs because reports on a drug's effects may be more likely if the effect is deleterious or severe. To address this issue, I included two 'control' drugs and followed the same procedure (i.e., literature review and extraction of outcomes). One 'control' drug was a known pregnancy safe drug – levothyroxine and the other 'control' drug was a known teratogenic drug – isotretinoin. I compared the findings

of drugs that modulate DHCR7 to these two 'controls' to determine if any effect I observe is significant and worthy of further investigation.

A certain number of adverse fetal outcomes occur in the general population. For additional comparison purposes, I also included data from the Centers of Disease Prevention and Control (CDC) on pregnancy outcomes (i.e., live born, spontaneous abortion, elective termination) collected across all races from both 1990 and 2008 for reference (CDC et al., 2012; Ventura et al., 2012). I coupled this with data stating that 3% of live-born babies (1 in 33) have a birth defect (CDC, 2008). **Table 20** displays the findings along with their corresponding references.

6.7.2.3 Fetal Outcomes from Prenatal Exposure to DHCR7 Modulators

Figure 29 contains the overall results for each of the 7 FDA approved DHCR7-modulating pharmaceuticals. For comparison purposes, known pregnancy-safe levothyroxine and known teratogen isotretinoin (or Accutane) are also included. The breakdown of these drugs across the 7 pregnancy outcomes is shown in **Figure 29A.** The number of deformations (possible teratogenic effects) vs. number of healthy babies is shown in **Figure 29B.** Drugs that inhibit DHCR7 resulted in deformities among 10.9% of babies born while drugs that increase DHCR7 expression resulted in deformities among 5.5%. This can be compared to the CDC background of 3.0% and levothyroxine's rate of 0.0% (the known pregnancy-safe drug). While DHCR7 inhibitors result in increased risk of deformities, the rate was still much lower than the well-known teratogen – Accutane or isotretinoin – at 47.7%. I also statistically compared how DHCR7 activity affected fetal outcomes. Using the CDC background rate from 2008 (CDC et al., 2012; Ventura et al., 2012) (accessed in January 2016), I performed a fisher's exact test for healthy vs. non-healthy (this includes born with birth defect/deformity, spontaneous abortion, elective terminations). As expected, levothyroxine (pregnancy safe drug) was not statistically

different from the background rate (p-value = 0.197). I found that DHCR7 inhibitors were highly enriched for adverse fetal outcomes (OR= 6.0, p-value < 0.001) with DHCR7 promoters showing less enrichment (OR=3.3, p-value < 0.001). Neither came close to the known teratogen isotretinoin (OR= 34.8, p-value < 0.001).

I also investigated the relationship between first-trimester exposure to DHCR7 modulators and increased risk of teratogenic effects (**Figure 29C** and **Figure 29D)**. First-trimester effects were especially interesting because reduction in cholesterol during the first-trimester has been shown to cause severe teratogenic effects (Edison and Muenke, 2004a). By inhibiting DHCR7 during the first-trimester, cholesterol production is also inhibited, therefore, I would expect severe teratogenic effects (Edison and Muenke, 2004a). Additionally, studies have shown that exposure to chemotherapeutic drugs, e.g., doxorubicin and tamoxifen, during the first-trimester is associated with increased risk of fetal complications (Germann et al., 2004). First-trimester results for all pregnancy outcomes are shown in **Figure 29C.** I grouped drugs by their DHCR7 effect in **Figure 29D.** Counts and percentages of pregnancy outcomes for first-trimester exposure to DHCR7 modulating drugs are provided in **Table 21.**

First-trimester exposure to drugs that increase DHCR7 expression (73.5% born healthy) was comparable to first-trimester exposure of a pregnancy-safe drug (72.7% born healthy). Even more importantly, first-trimester exposure to DHCR7 inhibitors was comparable to the known teratogen with 50.0% born healthy among drugs that inhibit DHCR7 vs. 48.3% for the teratogen. However, the sample size for first-trimester exposures was small (N=40 for known DHCR7 inhibitors) indicating that caution must be taken when interpreting these results.

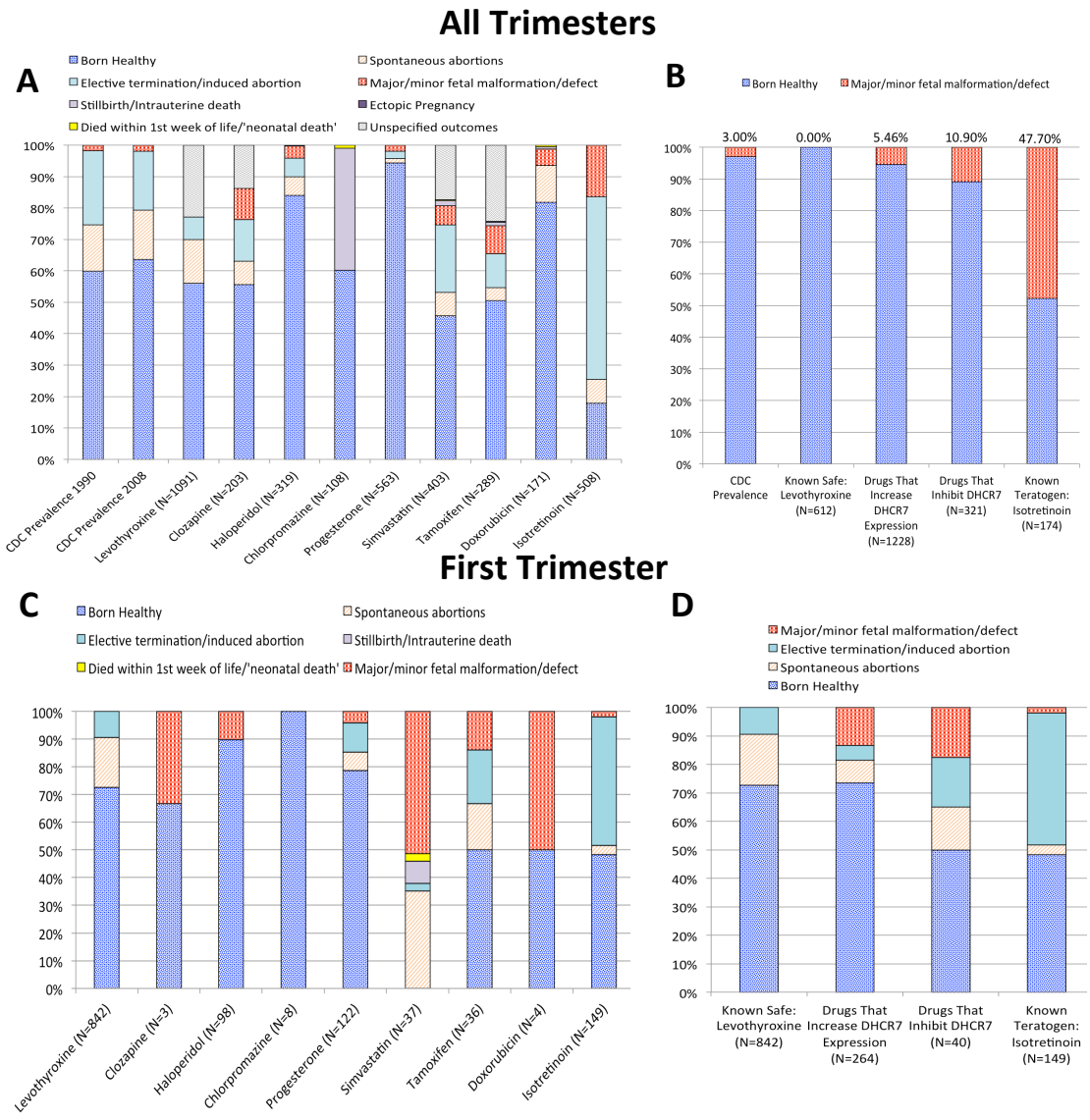


Figure 29. Fetal Outcomes of Prenatal Exposure to DHCR7 Modulators Compared to CDC Prevalence and a Known Teratogenic Drug (i.e., Isotretinoin or Accutane) and a Known Pregnancy Safe Drug (i.e., Levothyroxine). Figure 29A contains aggregated results across all trimesters for regarding pregnancy outcomes. **Figure 29B** contains only born healthy and born with fetal malformations, defects or congenital anomalies. Therefore, I excluded spontaneous abortions, elective terminations, stillbirths, ectopic pregnancies and neonatal deaths. Notice that in **Figure 29A** the number of fetal malformations/anomalies is higher among DHCR7 modulators including those that increase expression of DHCR7 (5.5%). However, drugs that inhibit DHCR7 have an increase in fetal malformations (10.9%). None of these levels comes close to the known teratogen, isotretinoin or Accutane, with 47.7% born with malformations (out of total born). Many patients that are pregnant elect to terminate their pregnancy (**Figure 29A),** which may be due to detected anomalies, however data on malformations among aborted fetuses is typically not available. I took all reported results where first-trimester exposure occurred (even if exposure persisted throughout the pregnancy) or where the exposure was listed as 'early pregnancy' as this appeared to indicate first-trimester exposure and these are shown in **Figure 29C**. Many known first-trimester Accutane or isotretinoin exposures resulted in elective terminations or induced abortions. Therefore, I included elective terminations and spontaneous abortions in **Figure 29D** along with live births (healthy or malformed).

Table 20. Reported Fetal Outcomes Following Prenatal Exposure to DHCR7 Modulating Drugs

Drug	Effect on DHCR7	Fetal Risk Category*	Fetal Pregnancy Outcome (N=No. of Patients)	Reference(s) on Fetal Outcome
Levothyroxine	None	A	Born Healthy (612) Spontaneous Abortion (151) Elective Termination (79) Fetal Malformation/Congenital Anomaly (0) Stillbirth/IntraUterine Death (0) Ectopic Pregnancy (0) Neonatal Death§ (0) Unspecified Outcome (249)	(Neto et al., 2007; Pomorski et al., 1999; Rotondi et al., 1999; Taylor et al., 2014; Zamperini et al., 2009)
Clozapine	Increases Expression	B	Born Healthy (113) Spontaneous Abortion (15) Elective Termination (27) Fetal Malformation/Congenital Anomaly (20) Stillbirth/IntraUterine Death (0) Ectopic Pregnancy (0) Neonatal Death§ (0) Unspecified Outcome (28)	(Gentile, 2004; Karakuła et al., 2004; McKenna et al., 2005; Reis and Källén, 2008; Stoner et al., 1997)
Progesterone	Increases Expression	B	Born Healthy (531) Spontaneous Abortion (8) Elective Termination (13) Fetal Malformation/Congenital Anomaly (11) Stillbirth/Intrauterine Death (0) Ectopic Pregnancy (0) Neonatal Death§ (0) Unspecified Outcome (0)	(Aarskog, 1979; Ahn et al., 2008; Hayles and Nolan, 1958)
Haloperidol	Increases Expression	C	Born Healthy (268) Spontaneous Abortion (19) Elective Termination (19) Fetal Malformation/Congenital Anomaly (12) Stillbirth/Intrauterine Death (0) Ectopic Pregnancy (1) Neonatal Death§ (0) Unspecified Outcome (0)	(Diav-Citrin et al., 2005; Mendhekar and Andrade, 2011; Newport et al., 2007; Reis and Källén, 2008; Yaris et al., 2004)
Chlorpromazine	Increases Expression	C	Born Healthy (65) Spontaneous Abortion (0) Elective Termination (0) Fetal Malformation/Congenital Anomaly (0) Stillbirth/Intrauterine Death (42) Ectopic Pregnancy (0) Neonatal Death§ (1) Unspecified Outcome (0)	(Chatterjee and Mukheree, 1997; Reis and Källén, 2008; Sawhney et al., 1998)
Simvastatin	Increases Expression	X	Born Healthy (184) Spontaneous Abortion (30) Elective Termination (87) Fetal Malformation/Congenital Anomaly (24) Stillbirth/Intrauterine Death (7) Ectopic Pregnancy (0) Neonatal Death§ (1) Unspecified Outcome (70)	(Edison and Muenke, 2004b; Manson et al., 1996; Pollack et al., 2005; Taguchi et al., 2008)
Tamoxifen (IC50=12nM; Kd=1nM)	Direct Inhibition	D	Born Healthy (146) Spontaneous Abortion (12) Elective Termination (31) Fetal Malformation/Congenital Anomaly (26)	(Andreadis et al., 2004; Barthelmes and Gateley, 2004; Berger and Clericuzio, 2008; Braems et al., 2011; Clark, 1993; Cullins et al., 1994; Isaacs et al., 2001; Koizumi

			Stillbirth/Intrauterine Death (3) Ectopic Pregnancy (1) Neonatal Death§ (0) Unspecified Outcome (70)	and Aono, 1986; Öksüzoglu and Güler, 2002; Tewari et al., 1997)
Doxorubicin (IC50=150-10000nM)	Direct Inhibition	D	Born Healthy (140) Spontaneous Abortion (20) Elective Termination (0) Fetal Malformation/Congenital Anomaly (9) Stillbirth/Intrauterine Death (1) Ectopic Pregnancy (0) Neonatal Death§ (1) Unspecified Outcome (0)	(Barni et al., 1992; Cardonick et al., 2012; de Wildt et al., 2009; Hahn et al., 2006; Karp et al., 1983; Kerr, 2005; Meyer-Wittkopf et al., 2001; Mir et al., 2012; Morris et al., 2009; Murray et al., 1984; Nieto et al., 2006; Peccatori et al., 2004; Potluri et al., 2006; Soliman et al., 2007; Willemse et al., 1990)
Isotretinoin	None	X	Born Healthy (91) Spontaneous Abortion (38) Elective Termination (296) Fetal Malformation/Congenital Anomaly (83) Stillbirth/Intrauterine Death (0) Ectopic Pregnancy (0) Neonatal Death§ (0) Unspecified Outcome (0)	(Dai et al., 1992; Honein et al., 2001; Schaefer et al., 2010)

* Fetal Risk Categories are based on Food Drug Administration's Criteria for the United States of America
§ Died within 1st week of life

Table 21. First-Trimester Fetal Outcomes of DHCR7 Modulating Drugs

N (%)	Known Safe: Levothyroxine	Drugs That Increase DHCR7 Expression	Drugs That Inhibit DHCR7	Known Teratogen: Isotretinoin
Healthy	612 (72.684)	194 (73.485)	20 (50.000)	72 (48.322)
Spontaneous abortion	151 (17.933)	21 (7.955)	6 (15.000)	5 (3.356)
Elective termination/ Induced abortion	79 (9.382)	14 (5.303)	7 (17.50)	69 (46.309)
Fetal malformation/ Defect	0 (0.000)	35 (13.258)	7 (17.50)	3 (2.013)
Overall Total	842	264	40	149

6.7.2.4 Interpretation of Fetal Results

I found that tamoxifen has an IC-50 of 12nM indicating that it inhibits DHCR7 at a pharmacologically potent level (Hopkins and Groom, 2002; Keller et al., 2006). I also found that tamoxifen exposure (especially during 1st trimester) resulted in increased fetal malformation risk. When compared to tamoxifen, doxorubicin had a lower ability to inhibit DHCR7 (IC-50 =150-10000nM) and likewise a lower risk of adverse fetal outcomes.

In total, five pregnancies resulted in fetal malformations/defects or congenital anomalies from first-trimester tamoxifen exposure while two pregnancies resulted from first-trimester doxorubicin exposure. The anomalies resulting from first-trimester tamoxifen exposure appeared to have phenotypes similar to SLOS. These included, two cases of genital deformities including ambiguous genitalia (Barthelmes and Gateley, 2004; Braems et al., 2011; Tewari et al., 1997), cleft palate (a form of holoprosencephaly) (Berger and Clericuzio, 2008; Braems et al., 2011) and Goldenhar's syndrome, which is characterized by facial deformities (Barthelmes and Gateley, 2004; Cullins et al., 1994), and one hand deformation (Braems et al., 2011). Note that facial deformities and cleft palate (a form of holoprosencephaly) can occur in SLOS patients. Genital anomalies occur frequently among SLOS patients (Fukazawa et al., 1992; Jira et al., 2003; Kelley and Hennekam, 2000). Hand deformations are rare among SLOS patients, however toe anomalies (e.g., toe syndactyly) are common (Jira et al., 2003; Kelley and Hennekam, 2000).

Two deformed babies reported in the literature were born after first-trimester doxorubicin exposure. One had genitalia issues (Murray et al., 1984) while the other had hydrocephalus (Potluri et al., 2006). These congenital defects are highly related to the pleiotropic effects of SLOS (and defective functioning of DHCR7) indicating the distinct possibility that

pharmacological inhibition of DHCR7 during the first-trimester of pregnancy results in teratogenic effects similar to the physical manifestations of SLOS.

6.7.2.5 Comment on Vitamin D3's In-Direct Inhibition of DHCR7

Of note, vitamin D (cholecalciferol) decreases the DHCR7 activity through an indirect mechanism (Zou and Porter). There has been some speculation on the possibility of teratogenic affects resulting from high doses of vitamin D when taken as a prenatal supplement (Roth, 2011). However, a recent review of vitamin D supplementation among pregnant women found that vitamin D supplementation may reduce the risk of pre-eclampsia, low birth weight, preterm birth (De-Regil et al., 2012; De-Regil et al., 2016) and adverse kidney outcomes in offspring (Miliku et al., 2015). Because the mechanism of vitamin D3's inhibition of DHCR7 is indirect (unlike the studied pharmacological DHCR7 inhibitors) it is likely to be mediated through a complex pathway with many interacting feedback loops. Studies involving moderate prenatal vitamin D supplementation have reported no adverse effects (Roth, 2011). Therefore, I would not expect vitamin D to result in SLOS-related teratogenic effects, although additional studies would be needed to ascertain the effects of large doses of prenatal vitamin D supplementation.

6.8 Limitations

There are several limitations for this literature review. First, information on what drugs target DHCR7 (either through inhibition or increasing gene expression) is limited by the current knowledge of drugs contained within Chembl. Not all drugs have been tested to determine the IC50 for its effect on DHCR7. Therefore it is possible that many more drugs inhibit DHCR7 as an off-target effect (i.e., not the main target of the drug). Second, the information on prenatal exposures to different compounds (both those that increase gene expression and those that inhibit the protein product) is limited by reported case studies in the literature. There is recall bias in

studies comparing the reported experiences of women during pregnancy versus after pregnancy (Bryant et al., 1989). These studies may also be afflicted with some of these biases. Further there may be under-reporting of miscarriage and fetal loss rates as these are often difficult to capture accurately in case studies.

6.9 Future Directions and Conclusion

In this review, I demonstrate the utility of an in-depth exploration of one Mendelian orphan disease: SLOS. I contribute a compendium of SLOS-inducing DHCR7 mutations, their incidence, and geographic distribution. I also include the incidence of these mutations in a large population from ExAC. This helped to illustrate important differences between mutation frequencies in SLOS vs. an assumed healthy population. It is believed that the ExAC population are non-SLOS individuals, however it is likely that they have other conditions. Comparing my SLOS mutation compendium to ExAC allele frequencies allows me to raise some important research questions both for studying under-represented ethnic groups, such as Africans (found to exhibit theoretically damaging DHCR7 mutations in ExAC), and for understanding genetic drift of DHCR7 mutations.

I went one step further and explored the prenatal effects of pharmaceuticals that target DHCR7. I posited that pharmaceutically induced DHCR7 inhibition would result in fetal outcomes similar to those seen in SLOS (miscarriage, fetal malformations, ambiguous genitalia). I reviewed the literature for observational studies on fetal outcomes due to drug exposure. I found that exposure to DHCR7 inhibitors during the first-trimester of pregnancy resulted in fetal deformities/malformations or anomalies in 50% of conceptions born healthy vs. 48% for a known teratogen control. Contrastingly, 73% were born healthy to those on drugs that increased DHCR7 expression compared to 73% for a known pregnancy-safe control. These results indicate

that screening for DHCR7 inhibition during the pre-clinical phase of drug toxicity may be helpful in identifying drugs with potential to induce adverse fetal outcomes before the drug is released to the market.

6.10 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in The Pharmacogenomics Journal (Boland and Tatonetti, 2016a) and presented at the American Society of Human Genetics in Vancouver, Canada (Boland and Tatonetti, 2016b). Support for this research provided by R01 GM107145 (MRB, NPT). MRB was supported by the National Library of Medicine training grant T15 LM00707 (MRB) from Jul 2014 – Jun. 2016 when this work was conducted.

Chapter 7

Development of a Machine Learning Approach to Classify Drugs Of Unknown Fetal Effect

7.1 Abstract

Many drugs commonly prescribed during pregnancy lack a fetal safety recommendation – called FDA 'category C' drugs. I present a data-driven approach to determine a drug's likelihood for inducing fetal harm, focusing on fetal loss and congenital anomalies. The fetal loss cohort contains 14,922 affected and 33,043 unaffected pregnancies and the congenital anomalies cohort contains 5,658 affected and 31,240 unaffected infants. I trained a random forest to classify drugs of unknown pregnancy class. Models achieved an out-of-bag accuracy of 91% for fetal loss and 87% for congenital anomalies outperforming null models. Fifty-seven 'category C' medications were classified as harmful for fetal loss and eleven for congenital anomalies. This includes medications with documented harmful effects, including naproxen, ibuprofen and rubella live vaccine. I also identified several novel drugs, e.g., haloperidol, that increased the risk of fetal

loss. My approach provides important information for pharmacologists and prescribers interested in drugs' fetal effects.

7.2 Introduction

In the late 1950s a great medical tragedy began with the promotion of thalidomide, an approved sedative, as a new modern treatment for morning sickness (Dally, 1998). Thousands of pregnant women began taking the drug, which resulted in a dramatic increase in spontaneous abortion (i.e., 'miscarriage'), and congenital abnormalities most notably shortened legs (Kim and Scialli, 2011). By mid-1961 it became clear that thalidomide was the culprit behind the increase in malformations leading to the drug's removal from the market (Smithells, 1962) and banned among women who may become pregnant. This experience led to the implementation of more stringent guidelines for drugs targeted at pregnant females.

Over the years the number of medications taken by pregnant women has grown. Concern over this 'epidemic of prescribing' among pregnant women began in the 1970s (Hill, 1973). A study of Danish women found that 44.2% of women received prescriptions for at least one medication during pregnancy (Olesen et al., 1999). Anti-inflammatory drugs are commonly prescribed medications in pregnancy, which have been shown to increase the risk of miscarriage or fetal loss (Nielsen et al., 2001). However, in many cases the effects that specific pharmacologics have on fetal outcomes remains unknown. The Food and Drug Administration (FDA) lists pharmacological drugs with unknown fetal outcomes as category C ('risk not ruled out') drugs while those with known teratogenic properties (such as thalidomide) are listed as category X ('contraindicated in pregnancy'). An estimated 37.8% of pregnant women receiving medications during pregnancy received an FDA category C drug (Andrade et al., 2004). Therefore, detailed study of these enigmatic drugs is greatly needed.

The purpose of this study is to systematically investigate the fetal outcomes, both fetal loss and congenital anomalies, following pharmacological exposure to category C drugs. This will provide both pharmacologists and physicians much needed information on the potential harms and benefits of these 'unknown fetal effect' drugs when they are considering prescribing them for pregnant women. Because fetal loss and congenital anomalies are two very different outcomes, I perform two separate cohort studies. In addition, I will focus some discussion on FDA category C drugs that target DHCR7, which are known to affect fetal outcomes (Boland and Tatonetti, 2016a).

7.3 Methods

7.3.1 Clinical Cohorts

7.3.1.1 Maternal Prescription Exposure and Fetal Outcome: Live Birth

Records were obtained on all infants born at the Columbia University Medical Center (CUMC) - New York Presbyterian Hospital (NYPH) healthcare system who had mothers listed in the Electronic Health Record (EHR) system. These links were created in the EHR system upon delivery to facilitate maternal-fetal care post-delivery. All mother-infant pairs with at least one medication prescribed before birth and up to 15 months prior were retained. I excluded all multiple infant pregnancies (e.g., twins, triplets), as these pregnancies are high-risk and subject to complications. I also excluded all pregnancies with any chromosomal abnormality diagnosed within the first three months of life (0-90 days of life). Presence of chromosomal abnormality was determined using the International Classification of Diseases, 9th edition (ICD-9) range 758-758.9.

Infants with congenital anomalies were identified as those having a congenital anomaly ICD-9 diagnosis, i.e., 740-759 (with 758-758.9 excluded) occurring within the first 90 days of life. Only

one anomaly diagnosis was necessary for identification although some infants had multiple anomalies. Minor anomalies were identified using criteria established set by the New York State Department of Health. Only minor anomalies within the 740-759 range were used (NYSDOH, 2007).

7.3.1.2 Maternal Prescription Exposure and Fetal Outcome: Fetal Loss

All pregnancies ending in fetal loss were identified as recorded at CUMC-NYPH. Fetal loss was defined as an ICD-9 code within the range: 630-639. Because I am interested in fetal outcomes following pharmacological exposure, I only included females with at least one medication prescribed up to 15 months prior to fetal loss. Fetal loss is this study includes spontaneous abortion (i.e. 'miscarriages'), legal/elective termination and any other forms of fetal loss/death recorded within the ICD-9 range 630-639. Because a female may have more than one fetal loss code occurring on two separate dates (often during the course of a single hospital visit), I collapsed dates to the month level. For the control population, I used women with a successful fetal outcome (i.e., single live birth) recorded at CUMC-NYPH with at least one medication prescribed up to 15 months prior to birth and who had no diagnosis of fetal loss recorded at the hospital and whose infant was without chromosomal abnormality.

7.3.2 Pharmacological Drug Information

The Food and Drug Administration (FDA) pregnancy categories for all drugs included in the study were extracted from uptodate.com and drugs.com. While the FDA has recently updated this labeling system and moved away from the A-X categorization schema (Boothby and Doering, 2001), I chose to use it in this study because it allows researchers and physicians to easily identify drugs with unknown fetal effects (i.e., the category C drugs). If a particular drug-combo was not listed with its own FDA pregnancy category designation then I used the most

severe pregnancy category from each drug in the combo. I also mapped each drug to its first-level Anatomical Therapeutic Chemical (ATC) classification system, which classifies drugs based on the effects of the drug's active ingredients on various organ systems. I also extracted the Mendelian genes either inhibited or affected (regardless of mechanism) for each drug using information from the Online Mendelian Inheritance in Man (OMIM) (URL: https://www.omim.org/). Because drugs targeting genes involved in vitamin processes may affect fetal risk (either protective or injurious), I also identified drugs that target at least one 'vitamin' gene as noted on DisGeNET – a disease-gene association network (URL: http://www.disgenet.org/).

The primary goal of this study is to find drugs that increase or decrease the risk of fetal loss following prenatal exposure. However, some prescription drugs / medications are used to induce legal termination or to treat subsequent conditions (e.g., hemorrhage, excessive bleeding, pain) and these drugs could bias the analyses. Therefore, I identified drugs given to women where the first prescription of the drug was the same day that the woman's legal termination was performed. I calculated the proportion of legal terminations where a given prescription drug was first prescribed out of those terminations where prescription information was available. All drugs with at least 2% frequency were labeled as 'drugs typically prescribed with legal termination'.

7.3.3 Statistical Analysis

7.3.3.1 Identifying Trimester of Drug Exposure

For pregnancies that resulted in a single live birth, I used the average gestation period (i.e., 38 weeks) as reported by the Centers for Disease Prevention and Control (CDC) (CDC, 2015). I then divided the 38-week pregnancy into three equal-sized periods (12.67 weeks each) as 'trimesters'. For pregnancies that resulted in fetal loss, I used the average time to fetal loss. CDC

reported that 91.6% of legal terminations occur within 13 weeks gestation with many other forms of fetal loss occurring prior to 13 weeks as well (CDC, 2017). Therefore, an exposure could have only occurred during the first trimester (i.e., one 12.67 week period). I also investigated two pre-conception periods (each 3 months in size) where exposures could occur both for the fetal loss and congenital anomaly cohorts. This was to investigate the presence or absence of a drug's pre-conception effect.

7.3.3.2 Classifying Category C Drugs Into Harmful and Non-Harmful Pregnancy Categories

I only investigated drugs having at least 50 pregnancies across all exposure periods (e.g., first trimester, second trimester) to minimize statistical anomalies due to low data. I also excluded all drugs classified as FDA pregnancy category N (i.e., Not Classified) or drugs that were 'Not Listed'.

A logistic regression model was constructed to predict a binary pregnancy category either 'Detrimental to Fetus – D or X' or 'Not Harmful to Fetus – A or B'. Three models were built – one model for fetal loss, a second for congenital anomalies and a third for minor congenital anomalies only. This allowed me to determine Odds Ratios (OR) and significance in a full model. The full model includes 29 features: one for each of 14 ATC classifications, 5 features indicating the number exposed during each trimester category (3 trimesters plus two 3-month periods for the pre-conception period), 5 features indicated the proportion of exposed with an anomaly per trimester category, 1 binary indicator variable for whether Mendelian genes were inhibited (from OMIM), 1 binary indicator variable whether Mendelian genes were affected (from OMIM), one binary indicator variable for whether vitamin genes are affected (from DisGeNET), one binary indicator variable for whether or not the drug could be used as a prenatal supplement (e.g., vitamin, mineral, glucose), and one binary indicator variable for whether or not

the drug was a treatment for nicotine abuse (since exposure to smoking during the prenatal period is a known risk factor for fetal loss and anomalies). For the fetal loss model, I only had 25 features because the majority of fetal losses occur during the first trimester and therefore I did not have variables for second and third trimester (either proportion of anomalies or exposed).

Next, a random forest model was constructed for both fetal loss and congenital anomalies (separately) with 2000 trees using all possible features. Out-Of-Bag (OOB) error rates were estimated to assess the quality of each model. Features were ranked using the Mean Decrease in Accuracy (MDA) with more informative features having higher MDAs. I compared a drug's probability of being harmful from each model for drugs with known FDA status and those with no recommendation (i.e., FDA category C drugs). Component analysis was performed using the proportion of fetal loss or the proportion with congenital anomaly per trimester of exposure to illustrate the relationship between adverse fetal outcomes and FDA pregnancy category. Code was implemented using R version 3.3.0.

7.4 Results

7.4.1 Clinical Cohort

Females with live-born births at NYPH were extracted where data on maternal drug exposure was captured in the EHR (i.e., the female had at least one prescription recorded in the EHR within a 1.3-year period prior to the child's birthdate). Infants with congenital anomalies were identified as those having a diagnosis within 90 days of life. The resulting dataset contained 31,240 pregnancies resulting in a live birth without a congenital anomaly and 5,658 pregnancies with a congenital anomaly (either major or minor). This indicated a baseline incidence of 15.33% for congenital anomalies. Of pregnancies with a recorded anomaly, 1,588 pregnancies had a minor anomaly. This cohort is referred to as the 'congenital anomaly' cohort while the cohort

containing the subset with minor anomalies is referred to as the 'minor congenital anomaly' cohort. Demographics of all pregnant females in both cohorts (i.e., congenital anomalies and fetal loss) that were included in this study are displayed in **Table 22.**

For the 'fetal loss' cohort, patients were extracted with any recorded fetal loss/death as indicated by a diagnosis within the ICD-9 range 630-639. For controls, patients with no prior fetal loss in their EHRs were selected having at least one single live birth recorded at NYPH. This resulted in a dataset of 14,922 pregnancies with fetal loss and 33,043 pregnancies without fetal loss and satisfying the above criteria. Fetal loss in this study includes any form of fetal loss/death recorded within the ICD-9 range 630-639.

7.4.2 Pharmacological Drug Dataset

The FDA pregnancy categories for all drugs included in this study were obtained from uptodate.com and drugs.com. The categories and their descriptions are provided in **Table 23** along with the number of distinct drugs belonging to each category in both cohorts. Drugs labeled as 'N: Not Classified' or 'Not Listed' were excluded from all analyses. The most popular category was category 'C: Risk Not Ruled Out' followed by the lower risk category 'B: No Risk in Other Studies'. Interestingly more distinct drugs belonged to category 'D: Positive Evidence of Risk' then category 'A: No Risk in Controlled Human Studies'. I extracted the ATC first-level categories for all distinct drugs included in the analysis. The most popular categories were 'A: Alimentary Tract and Metabolism' followed by 'D: Dermatologicals' indicating that a wide diversity of drugs belonging to these two categories are prescribed to pregnant females (**Table 24)**. I also extracted information on genes listed in OMIM that the drugs in the dataset target, along with drugs that target vitamin-related genes from DisGeNET. Drugs were identified that were commonly prescribed with legal termination at NYPH as these could bias the fetal loss

results. Two drugs identified were common drugs used in chemical abortions: Mifepristone (200 MG) and Misoprostol (0.2MG) (Schaff et al., 1999) and these were commonly prescribed at NYPH with legal termination (15.1% and 14.6% respectively).

7.4.3 Classifying FDA Category C Drugs As 'Harmful or 'Not-Known-To-Be-Harmful'

I constructed a logistic regression model with a binary outcome variable representing a non-harmful pregnancy classification (i.e., FDA pregnancy category A or B) versus a severe pregnancy classification (i.e., FDA pregnancy category D or X). For both congenital anomaly models (i.e., all anomalies, and minor anomalies), I added all possible features that could inform the model to predict the accurate FDA pregnancy category. The odds ratios along with their 95% confidence intervals (CIs) are shown in **Figure 30.**

In the fetal loss model, a drug belonging to the respiratory system category (ATC: R) increased the probability that the drug was either an A or B drug whereas, a drug belonging to the systemic hormonal preparations category (ATC: H) increased the probability that the drug was a harmful drug – either D or X. In the congenital anomalies model, drugs were more likely to be classified as A or B when a high proportion of anomalies was observed following exposure during the third trimester (t3).

A simple random forest model was built using only the proportion with anomaly (for the congenital anomaly cohort) or the proportion with fetal loss (for the fetal loss cohort) at each trimester of exposure. The model was run with 1000 trees and I constructed a multi-dimensional scaling (MDS) component plot to illustrate the separation among drugs that can be achieved using only the proportion with anomaly/fetal loss. Fifty-eight medications were classified as harmful and 206 safe in the fetal loss cohort. Eleven medications were classified as harmful and 181 safe in the congenital anomalies cohort. **Figure 31** shows the separation between the known-

harmful drugs (category D or X) in bright red from the not-known-to-be-harmful drugs in light blue (category A or B). The separation between the harmful and non-harmful drugs is most evident for the fetal loss cohort. I also separated out drugs that are used in legal termination to show where in the various plots those drugs appear (right-hand portion of **Figure 31**)**.** In most cases drugs prescribe during legal terminations are harmful categories (category D or X) or cluster with known harmful drugs.

A clear relationship was observed between the first MDS component and the proportion of women experiencing fetal loss following first trimester exposure to the drug (**Figure 32**). Because some of this effect could be due to legal termination, I separated out drugs empirically determined to be involved in legal termination procedures. This showed that the relationship was not fully due to drugs involved in legal termination. I also investigated the relationship between the proportions of women experiencing an offspring with a congenital anomaly following exposure across the three trimesters, but these results were not as stark.

Table 22. Demographics of Pregnant Females Included in Study

	Fetal Loss Dataset			Congenital Anomaly Dataset		
Demographic	**Without Fetal Loss (N=33043)**	**With Fetal Loss (N=14922)**	**P**	**Without Congenital Anomaly (N= 31240)**	**With Congenital Anomaly (N= 5658)**	**P**
Ethnicity						
Hispanic	13060 (39.5%)	6558 (43.9%)	0.215	12721 (40.7%)	2055 (36.3%)	0.123
Not-Reported as Hispanic	19983 (60.5%)	8364 (56.1%)		18519 (59.3%)	3603 (63.7%)	
Race			0.226			
Asian	877 (2.65%)	147 (0.99%)		755 (2.42%)	148 (2.62%)	0.076
Black	3131 (9.48%)	1448 (9.70%)		2871 (9.19%)	539 (9.53%)	
Indian	59 (0.18%)	5 (0.03%)		52 (0.17%)	10 (0.18%)	
Other	9594 (29.0%)	3962 (26.6%)		8858 (28.4%)	1776 (31.4%)	
Pacific Islander	123 (0.37%)	87 (0.58%)		131 (0.42%)	24 (0.42%)	
Unidentified/Declined/Unknown	8580 (26.0%)	5775 (38.7%)		8672 (27.8%)	1634 (28.9%)	
White	10679 (32.3%)	3498 (23.4%)		9901 (31.7%)	1527 (27.0%)	
Age at birth/fetal loss (median and first-third quartile)	29.28 (24.03-34.40)	27.54 (22.76-33.46)	<0.001	28.92 (23.80-34.14)	29.27 (24.15-34.25)	0.382

Table 23. FDA Pregnancy Categories and Descriptions

FDA Category	Description	Studies In Humans	Studies In Animals	No. of Drugs in Fetal Loss Dataset (N=499)*	No. of Drugs in Congenital Anomaly Dataset (N=378)*
A	No Risk in Controlled Human Studies	No Risk To Fetus	No Risk To Fetus	20	15
B	No Risk in Other Studies	No Adequate Studies OR No Risk To Fetus	No Risk To Fetus OR **Risk To Fetus**	147	125
C	Risk Not Ruled Out	No Adequate Studies	**Risk To Fetus**	264	192
D	Positive Evidence of Risk	**Risk To Fetus**	**Risk To Fetus**	37	26
X	Contraindicated in Pregnancy	**Risk To Fetus** – Including Anomalies	**Risk To Fetus**	31	20
N	Not Yet Classified Into A Pregnancy Category	-	-	-	
Not Listed		-	-	-	

***Distinct Drug-Dosage Combos – A drug only has 1 FDA Pregnancy Classification**

Table 24. ATC Classifications and Descriptions

ATC Category	Description	No. of Drugs in Fetal Loss Dataset*	No. of Drugs in Congenital Anomaly Dataset*
A	Alimentary Tract and Metabolism	131	113
B	Blood and Blood-Forming Organs	44	38
C	Cardiovascular System	53	41
D	Dermatologicals	81	65
G	Genito-Urinary System and Sex Hormones	81	61
H	Systemic Hormonal Preparations, excluding sex hormones and insulins	21	18
J	Anti-infectives for Systemic Use	69	58
L	Anti-neoplastic and Immuno-modulating agents	15	10
M	Musculo-skeletal System	17	11
N	Nervous system	72	44
P	Anti-parasitic products, insecticides and repellents	6	6
R	Respiratory System	76	53
S	Sensory Organs	60	49
V	Various	12	12

***Distinct Drug-Dosage Combos – A drug can have multiple ATC classifications**

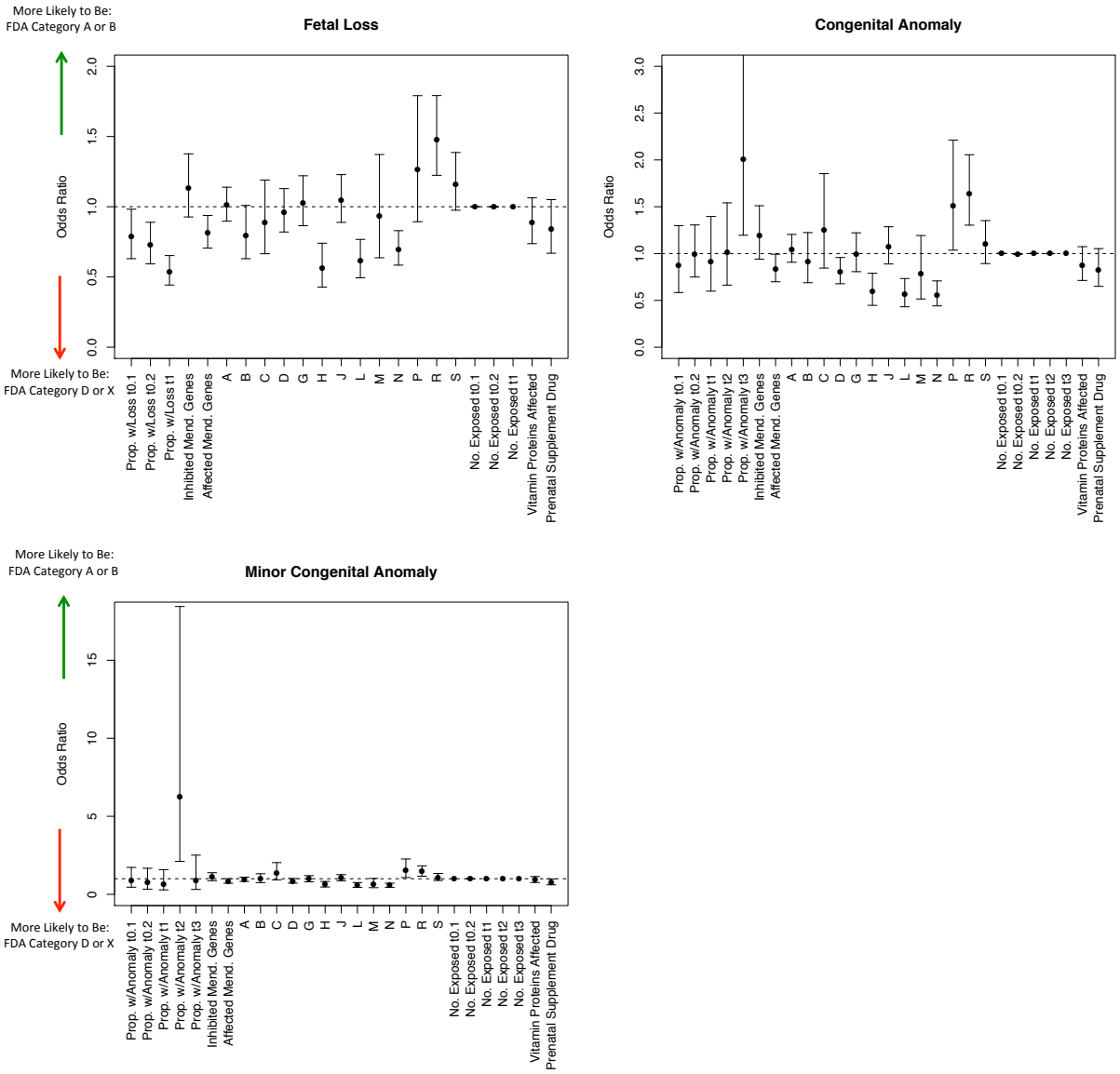


Figure 30. Odds Ratios from Logistic Regression Models: Fetal Loss, Congenital Anomaly and Minor Congenital Anomaly. ORs less than 1 indicate that the model predicts that feature to indicate that a drug is more likely to be a FDA Category D or X Drug. ORs greater than 1 indicate that the model predicts that feature to indicate that a drug is more likely to be a FDA Category A or B drug. In the fetal loss model, a drug having an ATC category of R (Respiratory System drug) increases the probability that the drug is an FDA Category A or B drug (i.e., a 'good' drug) while an ATC category of H (Systemic Hormonal Preparations excluding sex hormones and insulins) increases the probability that the drug is FDA Category D or X (i.e., a 'harmful' drug).

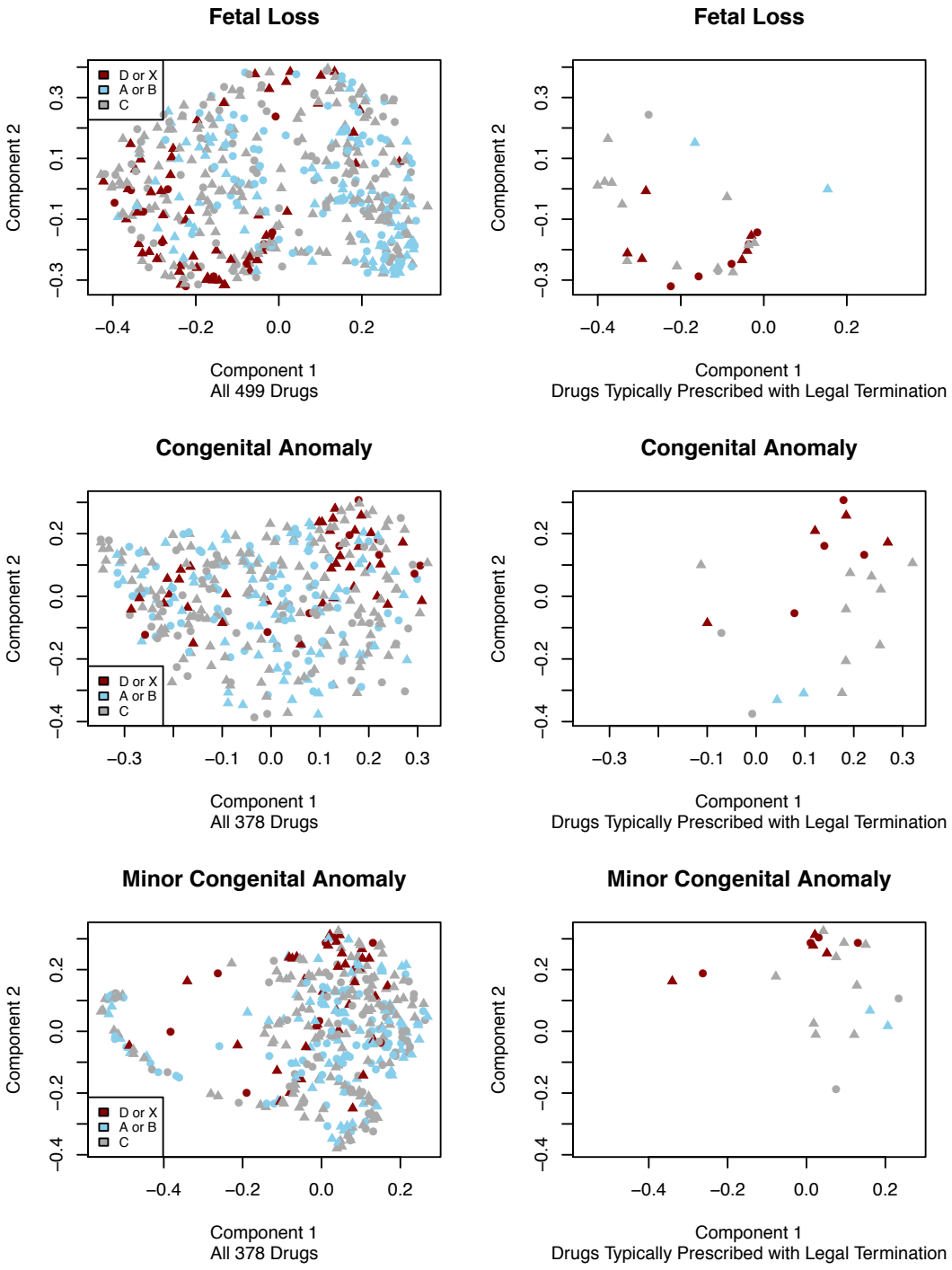


Figure 31. Multi-Dimensional Scaling (MDS) Component Plots for: Fetal Loss, Congenital Anomaly and Minor Congenital Anomaly. The three subplots on the left hand side of the figure contain all drugs while on the right hand side of the figure contain only drugs typically prescribed with legal termination. Red drugs are those shown as category D or X, blue drugs are category A or B while category C drugs are shown as grey. For congenital anomalies, the proportion with an anomaly for each of the 5 exposure periods (2 pre-conception and 3 trimesters) were included as features in this principal component analysis. For fetal loss, only the 2 pre-conception periods and the first trimester were included.

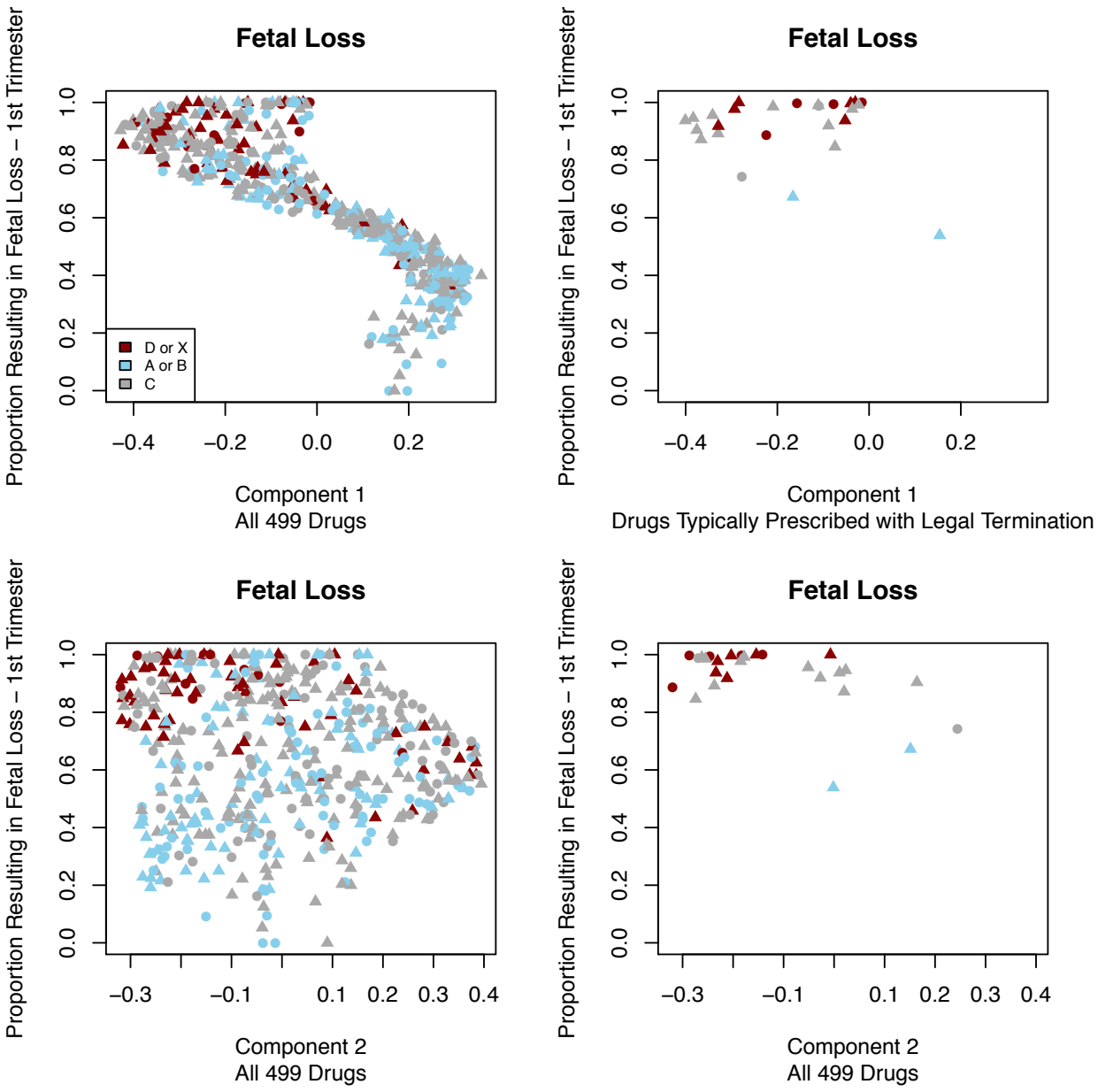


Figure 32. Component vs. Proportion with Fetal Loss. There is a clear relationship between the first component and the proportion of individuals experiencing fetal loss following prenatal exposure to the drug. This effect is not entirely due to drugs prescribed for legal termination, which are shown separately in the right most subplots.

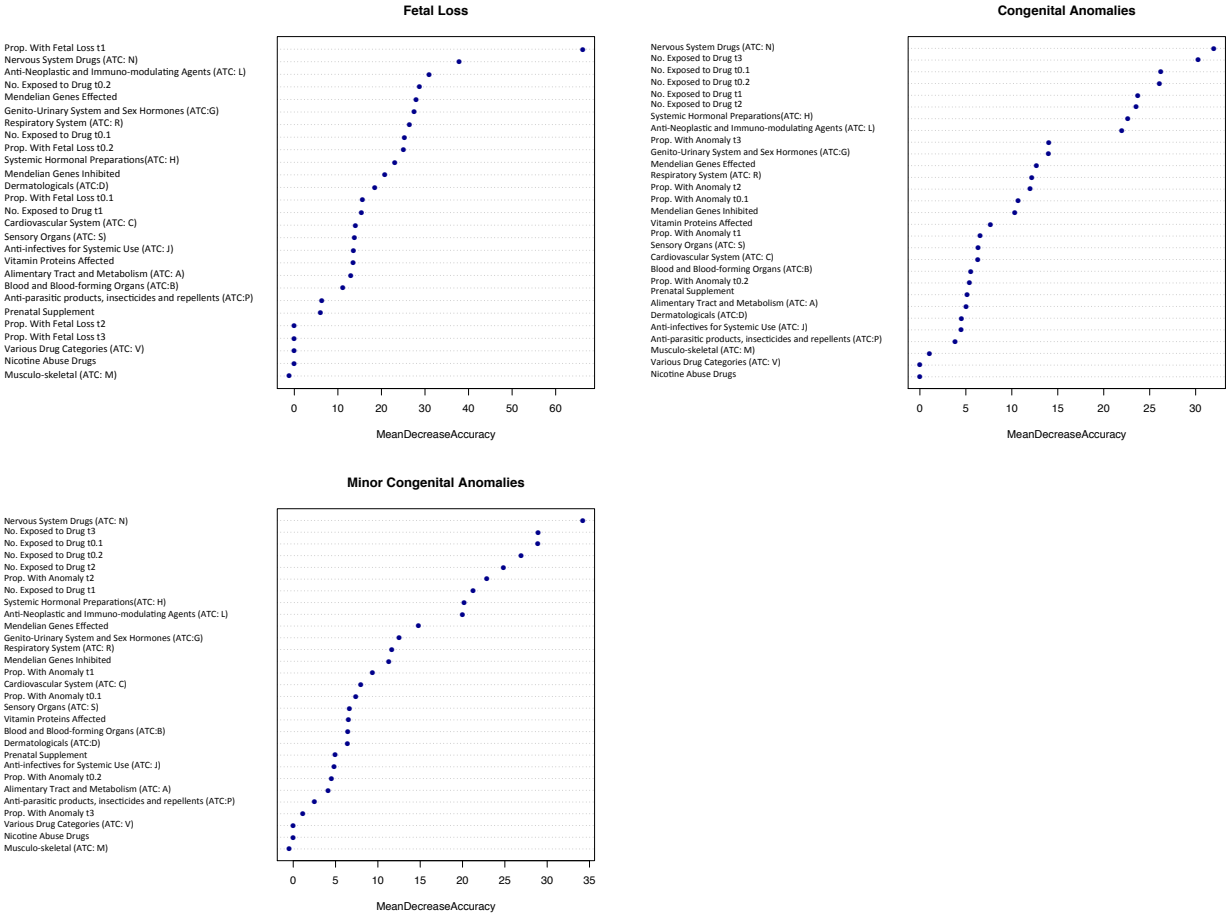


Figure 33. Mean Decrease in Accuracy (MDA) Plots for: Fetal Loss, Congenital Anomaly and Minor Congenital Anomaly. The higher the MDA score, the more informative the feature is to the model's performance. This means that features towards the top of each subplot are the most indicative of whether a drug is known to be harmful (D or X) versus not-known-to-be-harmful (A or B). The number of individuals exposed to a drug at each trimester was highly informative of the drug's pregnancy class (i.e., harmful (D or X) or not-known-to-be-harmful (A or B)). This is intuitive as physicians often modify their behavior when they identify a woman as being pregnant and are less likely to give them a known harmful drug (i.e., D or X). ATC drug class N is also very predictive in all 3 models, but was more predictive for the congenital anomalies models then the fetal loss model.

Next I ran the model containing all potentially informative features in the random forest model with 2000 trees. Each feature's contribution to the model's performance was assessed using Mean Decrease in Accuracy (MDA). Features with high MDA are more important in contributing to the model's performance. I found that the number of individuals exposed to a drug at a given trimester was highly informative in predicting whether a drug was harmful (i.e., D or X) versus not-known-to-be-harmful (A or B) as shown in **Figure 33**. This is intuitive as physicians often changing their prescribing behavior upon recognizing that a woman is pregnant. Therefore, they are not as likely to prescribe a new medication that is known to be harmful to a woman who is pregnant (D or X drug) making this feature to be very informative in all three outcome models. Interestingly, the proportion of those born with an anomaly following exposure to a drug for a given trimester was not as informative in the model indicating that FDA class affects the exposure pattern, but is not necessarily based on the overall fetal toxicity (as this is often unknown at the time the FDA class is assigned).

Certain ATC drug classes were also found to be very informative in the model (**Figure 33)**. The top five most informative ATC classes in predicting a drug's FDA pregnancy category across all fetal outcome models were nervous system drugs (ATC: N), systemic hormonal preparations excluding sex hormones (ATC: H), anti-neoplastic and immune-modulating agents (ATC: L), genito-urinary system and sex hormones (ATC: G) and respiratory system (ATC: R). The ordering of the specific categories importance varied by model with ATC category N being the most informative feature overall in both the congenital anomalies (major and minor) and the minor anomalies only models.

Importantly, a binary indicator variable for whether or not vitamin genes (from DisGeNET) were affected by a given drug was consistently more informative then whether or not a drug was

actually a prenatal supplement/mineral/vitamin (**Figure 33**). Additionally, whether or not Mendelian genes were affected or inhibited was more informative then whether or not a vitamin gene was affected indicating the importance of Mendelian genes in developmental processes.

The out-of-bag (OOB) estimated error rate was 9.36% for the fetal loss model (containing 235 / 499 drugs with known non-C FDA class), and 12.90% for both the congenital anomalies model (containing 186 / 378 drugs with known non-C FDA class) and the minor anomalies model (also containing 186 / 378 drugs with known non-C FDA class). This indicates that the estimated accuracy was 90.64% for the fetal loss model, and 87.10% for both anomalies models.

Drugs predicted by the models to be similar to harmful drugs (D or X) in the fetal loss model are displayed in **Table 25** (Overall OOB accuracy: 90.64%) and **Table 26** shows drugs predicted to be harmful in the congenital anomaly model (Overall OOB accuracy: 87.10%). All known harmful drugs (D or X) had a model probability$_{harmful}$ above 50% while all not-known-to-be-harmful (or safe) drugs (A or B) had a model probability$_{harmful}$ below 50% (**Figure 34**). Category C drugs, where no recommendation for pregnancy status is given, had probabilities of being harmful across the entire spectrum (**Figure 34**). All 192 category C drugs included in the congenital anomalies model were also included in the fetal loss model (fetal loss model contained 264 category C drugs). This allowed for the comparison of the probability that a drug was harmful in increasing the risk of fetal loss and congenital anomalies. These two probabilities were highly correlated ($r=0.63$, $p<0.001$). Drugs like rubella virus vaccine were predicted harmful in increasing the risk of fetal loss and also congenital anomalies (**Figure 35**). Some drugs, like Fentanyl and Benzocaine were only predicted harmful in one model. These drugs are likely to require further investigation to determine if there is a mechanistic reason for this difference.

Two drugs were predicted by the congenital anomalies model to be similar to harmful drugs that had very few exposures during the entire pregnancy (first through third trimesters) these drugs are Hydromorphone Hydrochloride (2MG) and Benzocaine (200MG/ML mucosal spray). Typically, physicians reduce pregnant females exposure to drugs that are labeled as category D or X drugs resulting in very few exposures during the first through third trimesters. The model detects drugs based on similar patterns. Since these two drugs had very low exposures during pregnancy they were detected as being potentially harmful drugs because their pattern of usage was similar. Other drugs, such as naproxen (2 different dosages) had a high rate of anomalies among exposed infants (even though the exposures were reduced during pregnancy versus pre-conception period), which was another hallmark of category D or X drugs. The advantage of a machine learning approach is the ability to detect these patterns in prescribing (i.e., number of exposed) and also fetal severity of the drug (i.e., proportion with anomalies) in determining the predicted pregnancy class for category C drugs.

7.5 Discussion

The models successfully identified category C drugs that are likely to be harmful (D or X) and not-likely-to-be-harmful (A or B) across either fetal loss or congenital anomalies. This information is critical for pharmacologists seeking to understand placental and trans-placental effects and for physicians who may be considering prescribing a category C drug to a pregnant female.

Table 25. Category C Drugs Predicted to be Harmful (D or X): Fetal Loss Cohort

Drug Name	Percent With Fetal Loss t_{-2}	Percent With Fetal Loss t_{-1}	Percent With Fetal Loss t_1
Alimentary Tract and Metabolism (ATC: A)			
Sodium Chloride 0.0769 MEQ/ML Injectable Solution	36.8	50	73.9
3 ML Sodium Chloride 9 MG/ML Prefilled Syringe	68.8	90.5	74.1
Calcium Chloride 0.001 MEQ/ML / Glucose 50 MG/ML / Potassium Chloride 0.004 MEQ/ML / Sodium Chloride 0.103 MEQ/ML / Sodium Lactate 0.028 MEQ/ML Injectable Solution	39.8	57	87.4
Magnesium Oxide 400 MG Oral Tablet	58.3	50	81.3
Calcium Gluconate 100 MG/ML Injectable Solution	54.5	52.2	82.5
Potassium Chloride 0.4 MEQ/ML Injectable Solution	41.7	46.7	88.1
Potassium Chloride 10 MEQ Extended Release Oral Tablet	36.4	55.6	86.4
Magnesium Hydroxide 80 MG/ML Oral Suspension	38.8	53.5	92.3
Calcium Chloride 0.0014 MEQ/ML / Potassium Chloride 0.004 MEQ/ML / Sodium Chloride 0.103 MEQ/ML / Sodium Lactate 0.028 MEQ/ML Injectable Solution	43.6	47.6	87.2
Cardiovascular System (ATC: C)			
Dexamethasone 4 MG Oral Tablet	50	81.8	91.3
Hydrocortisone 25 MG/ML Topical Cream	50	72.7	70
Ibuprofen 800 MG Oral Tablet	59.3	71.8	99
Nifedipine 10 MG Oral Capsule	36.7	72.7	96.2
24 HR Nifedipine 30 MG Extended Release Oral Tablet	36.7	60	86.2
24 HR Nifedipine 60 MG Extended Release Oral Tablet	50	40	89.5
Furosemide 20 MG Oral Tablet	50	33.3	85.2
Labetalol hydrochloride 5 MG/ML Injectable Solution	23.5	28.6	88.5
Genitourinary System and Sex Hormones (ATC: G)			
Naproxen 500 MG Delayed Release Oral Tablet	47.8	43.8	84.6
Naproxen 500 MG Oral Tablet	46	49.1	83.3
Dinoprostone 10 MG Drug Implant [Cervidil]	43.9	73.1	100
1 ML Carboprost 0.25 MG/ML Injection [Hemabate]	56.3	85.7	100
Methylergonovine Maleate 0.2 MG Oral Tablet	53.3	75	99.1
Methylergonovine Maleate 0.2 MG Oral Tablet [Methergine]	45.9	75	100
Methylergonovine Maleate 0.2 MG/ML Injectable Solution [Methergine]	45.2	57.1	97.9
Pain-Reliever Combination Drugs (ATC: N)			
Acetaminophen 300 MG / Hydrocodone Bitartrate 10 MG Oral Tablet	77.8	100	99.2
Acetaminophen 325 MG / Codeine Phosphate 30 MG Oral Capsule	51.9	60.9	98.6
Acetaminophen 325 MG / Oxycodone Hydrochloride 5 MG Oral Tablet	36.6	45.3	90.4
Acetaminophen 650 MG Rectal Suppository [Acephen]	60.3	61.4	85.8
Anti-Depressant or Anti-Psychotic (ATC: N)			
Sertraline 100 MG Oral Tablet	60	41.7	68.4
Sertraline 25 MG Oral Tablet	71.4	50	75
Sertraline 50 MG Oral Tablet	53.3	41.7	76.2
Prochlorperazine 10 MG Oral Tablet	76.9	85.7	73.7
Prochlorperazine 5 MG Oral Tablet	75	64.3	84.6
Citalopram 20 MG Oral Tablet	27.3	29.4	71.4
Haloperidol 5 MG Oral Tablet	40	70	78.3
Haloperidol 5 MG/ML Injectable Solution	22.2	50	78.6
Fluoxetine 20 MG Oral Capsule	41.7	50	91.7
Trazodone Hydrochloride 50 MG Oral Tablet	57.1	38.9	45
Migraines or Anti-Seizure (ATC: N)			
Sumatriptan 25 MG Oral Tablet	62.5	65	75
Gabapentin 300 MG Oral Capsule	40	44.4	57.1
Sedative (ATC: N)			

Zolpidem tartrate 10 MG Oral Tablet	37.5	62.5	84.1
Zolpidem tartrate 5 MG Oral Tablet	39.8	49.8	90.3
Opioid (ATC: N)			
Butorphanol Tartrate 2 MG/ML Injectable Solution	62.2	75	83.3
Hydromorphone Hydrochloride 2 MG Oral Tablet [Dilaudid]	45	53.8	69.6
Hydromorphone Hydrochloride 2 MG/ML Injectable Solution	55.4	48.6	88
1 ML Hydromorphone Hydrochloride 1 MG/ML Injection	47.1	42.5	91.6
Morphine Sulfate 10 MG/ML Injectable Solution	14.3	61.5	90
Morphine Sulfate 2 MG/ML Injectable Solution	41.7	48.5	84.9
Morphine Sulfate 4 MG/ML Injectable Solution	50.9	54.7	90.9
12 HR Oxycodone Hydrochloride 10 MG Extended Release Oral Tablet	27.3	50	100
Oxycodone Hydrochloride 5 MG Oral Tablet	57.4	49.5	85
Tramadol hydrochloride 50 MG Oral Tablet	28.6	33.3	75
Fentanyl 0.05 MG/ML Injectable Solution	70.6	62.1	98.7
Respiratory System (ATC: R)			
Promethazine Hydrochloride 25 MG Oral Tablet	96.2	89.4	97.7
Musculoskeletal / Sensory System (ATC: M and S)			
1 ML Ketorolac Tromethamine 30 MG/ML Prefilled Syringe	46.4	47.9	91
Various Systems (ATC: V)			
Naloxone Hydrochloride 0.4 MG/ML Injectable Solution	55.6	57.1	96.1
Vaccine			
Rubella Virus Vaccine Live (Wistar RA 27-3 Strain) 2000 UNT/ML Injectable Solution	37	47.5	96.4

Bold italics indicates that drug was implicated in adverse fetal outcomes for both the fetal loss model and the congenital anomaly model.

t_{-2}: Pre-conception effect: -6 to -3 months before conception

t_{-1}: Pre-conception effect: -3 to 0 months before conception

t_1: First Trimester

Table 26. Category C Drugs Predicted to be Harmful (D or X): Congenital Anomalies Cohort

Drug Name	Percent With Anomaly t_{-2}	Percent With Anomaly t_{-1}	Percent With Anomaly t_1	Percent With Anomaly t_2	Percent With Anomaly t_3	Period At Risk
Pain Reliever						
Benzocaine 200 MG/ML Mucosal Spray	13.4	10.7	0	0	0	RGDP
NSAID						
Ibuprofen 200 MG Oral Tablet	12.2	18.6	29.2	0	0	1st
Ibuprofen 800 MG Oral Tablet	16.9	14.1	20	12.5	0	1st -2nd
Naproxen 250 MG Oral Tablet	14.3	8	50	0	0	1st
Naproxen 500 MG Oral Tablet	12.9	13.2	7.7	20	0	1st -2nd
Opioid						
Butorphanol Tartrate 2 MG/ML Injectable Solution	5	14.3	0	0	10.4	3rd
Hydromorphone Hydrochloride 2 MG Oral Tablet [Dilaudid]	9.1	14.3	0	0	0	RGDP
Hydromorphone Hydrochloride 2 MG/ML Injectable Solution	10.7	9.5	0	6.1	12.8	2nd – 3rd
Steroid						
Hydrocortisone 25 MG/ML Topical Cream	57.1	12.5	0	5.6	3.9	2nd – 3rd
Cervical Implant						
Dinoprostone 10 MG Drug Implant [Cervidil]	9.2	16.7	0	0	13.5	3rd
Vaccine						
Rubella Virus Vaccine Live (Wistar RA 27-3 Strain) 2000 UNT/ML Injectable Solution	12.6	8.5	20	0	15	1st or 3rd

***** RGDP: Rarely Given During Pregnancies Ending in Liveborn Infants. The dramatic drop off in prescribing throughout the entire pregnancy (t1-t3) is why these drugs were labeled as likely to be harmful during pregnancy (D or X) due to similar patterns being observed for drugs that are known to be harmful.

Bold italics indicates that drug was implicated in adverse fetal outcomes for both the fetal loss model and the congenital anomaly model.

t_{-2}: Pre-conception effect: -6 to -3 months before conception
t_{-1}: Pre-conception effect: -3 to 0 months before conception
t_1: First Trimester
t_2: Second Trimester
t_3: Third Trimester

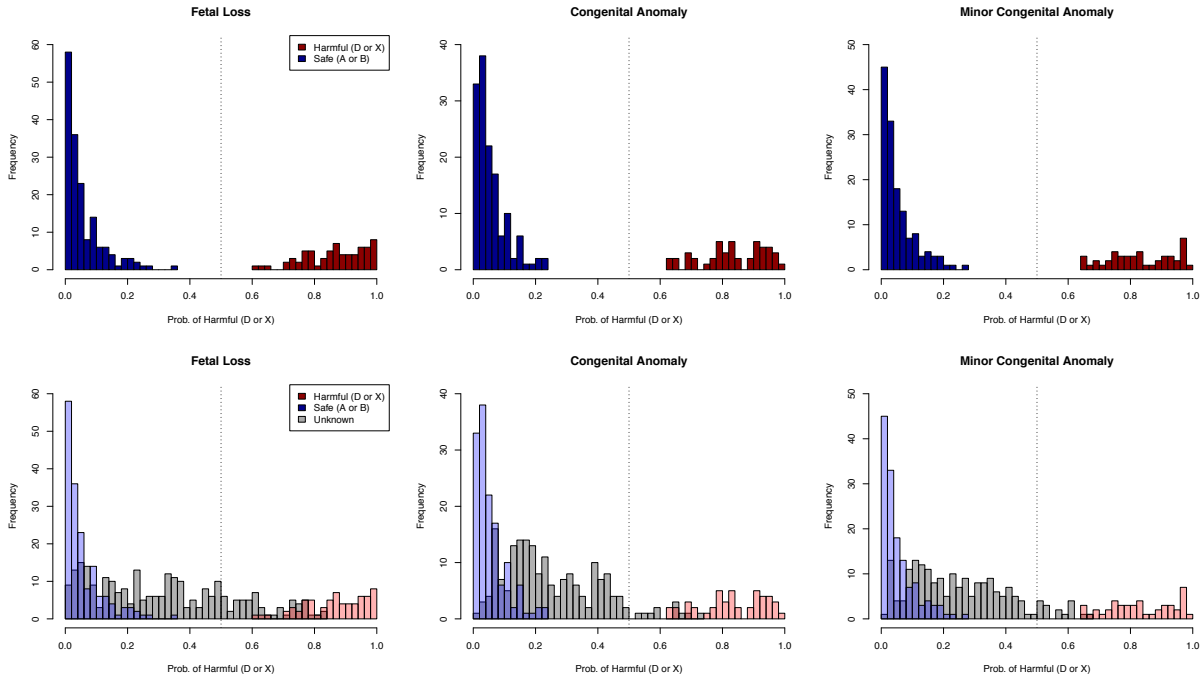


Figure 34. Model Probability of Being a Harmful Drug (D or X). The top portion of the graph shows drugs with known FDA pregnancy class. All drugs above the 50% probability threshold were predicted to be harmful and were harmful (three top graphs) across all models including fetal loss, congenital anomaly and minor congenital anomalies alone. In the lower three graphs FDA category C drugs are included (depicted in light grey). These drugs have no FDA recommendation regarding their safety during pregnancy. The majority of these drugs were predicted to be pregnancy safe (less than 50% probability of being harmful). While some drugs were above the 50% threshold and were more similar to known harmful drugs.

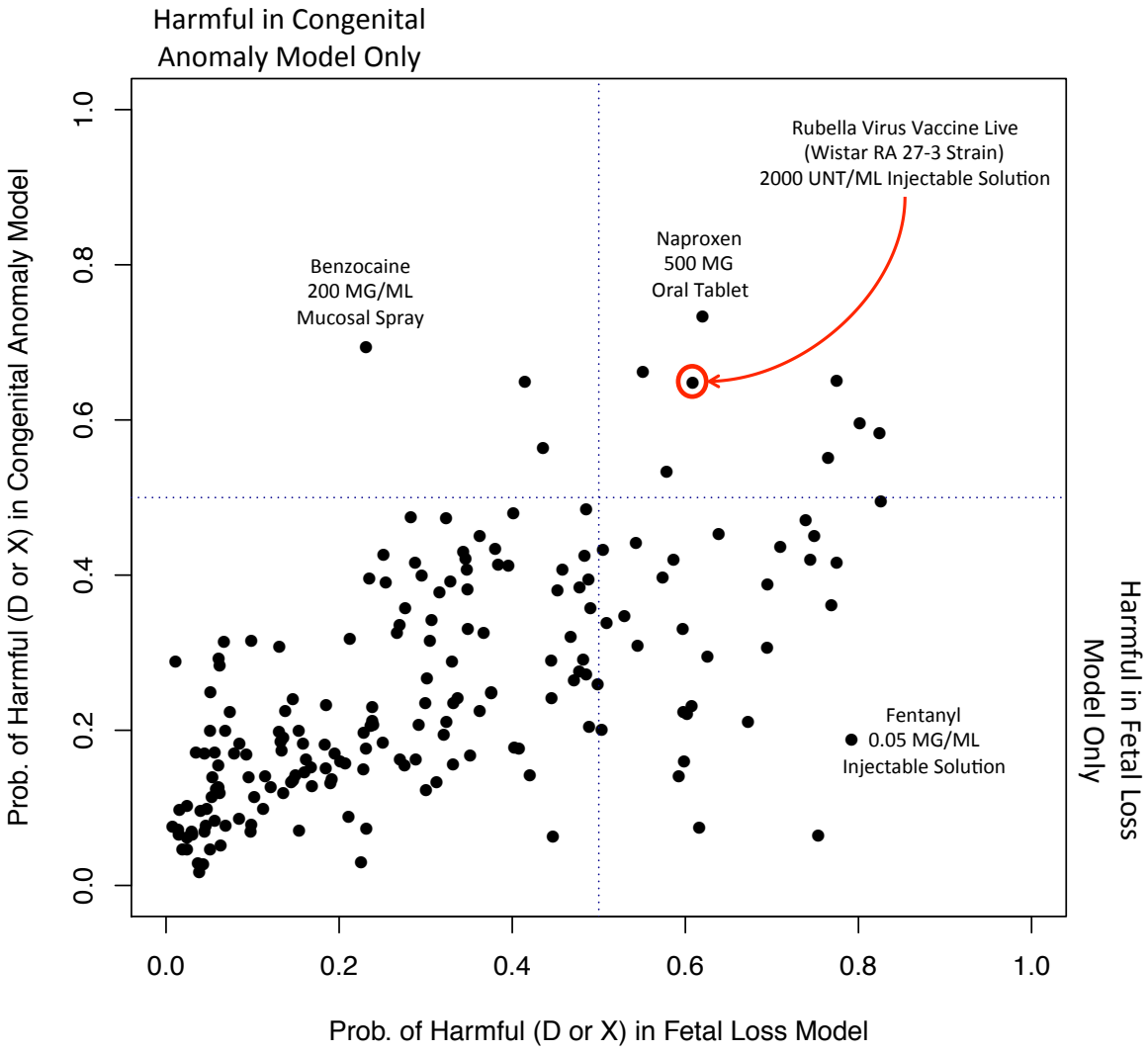


Figure 35. Model Probability of Being a Harmful Drug (D or X) in Congenital Anomaly Model vs. Fetal Loss Model for Category C Drugs (i.e., those with no FDA recommendation). The model probabilities for a drug's harmful status were highly correlated (r=0.63, p<0.001) between both congenital anomaly and fetal loss models. NSAIDs like naproxen were predicted harmful by both models. Also live rubella vaccination was harmful in both models. Other drugs were predicted harmful in increasing the risk of either fetal loss only (lower right hand quadrant) or congenital anomalies only (upper left hand quadrant). These may require further investigation to determine the mechanistic rationale for their predicted harm in one fetal outcome over the other.

7.5.1 Drugs Predicted Harmful in Congenital Anomaly Model

Eleven distinct medications or eight distinct drugs were predicted to be harmful (D or X) in the congenital anomalies model. I developed a machine-learning algorithm to predict drugs that were harmful based on anomaly rates and usage patterns for drugs with known FDA pregnancy classifications. This machine learning approach predicts a drug to be harmful if one of the following conditions is met: a.) drug exposure results in a high proportion of anomalies; b.) the drug's usage was greatly restricted during pregnancy (i.e., females were exposed during pre-conception period at much higher rates then during pregnancy; and c.) drug was similar to known harmful drugs in terms of mechanism (e.g., ATC classification, drug targets proteins involved in Mendelian diseases, drug targets known vitamin-related proteins).

7.5.1.1 Non-Steroidal Anti-Inflammatory Drugs (NSAIDs)

Two predicted harmful drugs (and 4 distinct medications) were Non-Steroidal Anti-Inflammatory Drugs (NSAIDs), namely ibuprofen and naproxen. Several studies report an increased risk of anomalies, specifically cardiac anomalies among infants exposed to either naproxen, ibuprofen, NSAIDs generally or combinations of those drugs (Ericson and Källén, 2001; Ofori et al., 2006). In most cases studying the effects of NSAIDs on fetal anomalies, the drugs - naproxen and/or ibuprofen - were often associated with the most number of congenital anomalies (Ericson and Källén, 2001; Hernandez et al., 2012). For both drugs, this study observed the highest risk among first and second trimester exposures, with higher risk among naproxen users than ibuprofen users (**Table 26).**

7.5.1.2 Live Rubella Vaccine

Maternal exposure to 'Rubella Virus Vaccine Live' was predicted to be harmful by my algorithm (D or X). It should be noted that live rubella vaccine is not indicated in pregnancy and often

occurred in the pre-conception and first-trimester period of the pregnancy. The proportion of females exposed to the vaccine whose infant was born with a congenital anomaly was high. For the pre-conception period 6-3 months prior to conception the rate of anomaly following exposure to live rubella vaccine was 12.58%, and 8.47% in the period 3-0 months prior to conception and a rate of 20% with anomalies for first trimester exposure. A very low number of females were exposed during the second trimester (most likely because live rubella vaccine is contra-indicated during pregnancy) and therefore no anomalies were reported among those first exposed during the second trimester. However, a rate of 15% with anomalies was observed for those exposed during the third trimester (**Table 26**).

The relationship between rubella virus and fetal anomalies is well known and described in the literature (Cooper and Krugman, 1967) (Swan et al., 1943; Webster, 1998). In fact, a single epidemic of rubella caused more birth defects in the United States then thalidomide during the entire time that it was on the global market (Webster, 1998). Increases in both anomalies and fetal loss were observed in women infected with rubella during pregnancy (Naeye and Blanc, 1965; Rudolph et al., 1965). Exposure to live rubella vaccine was considered harmful (D or X) in both the congenital anomaly models and also the fetal loss model (**Table 25**) indicating an increased risk of fetal loss as well as an increased risk of anomalies. This is consistent with the literature on the harms of inadvertent exposure to rubella during the early stages of pregnancy. This study also shows some evidence of a pre-conception effect at least 6 months prior to conception. Pre-conception effects of maternal rubella have been reported in past epidemics (Wolff, 1972), but not in the vaccination context. One study from Turkey found seventeen infants inadvertently exposed to the rubella vaccine either in the first month of pregnancy or just prior to conception and all 17 showed no signs of congenital rubella syndrome (Ergenoglu et al.,

2012). Evidence from the epidemics shows that third trimester exposure to rubella was associated with more ocular issues while preconception rubella was associated with severe outcomes, but no reported ocular issues (five pregnancies reported in literature: two spontaneous abortions, two neonatal deaths, and one live-born without ocular issues) (Wolff, 1972). From the fetal loss cohort, 96.4% of those receiving rubella vaccination in the first trimester (83 exposed during first-trimester in fetal loss cohort) resulted in a fetal loss (**Table 25**). This indicates the severity of first-trimester rubella exposure on fetal outcomes underscoring the importance of avoiding rubella vaccination prior to conception.

7.5.1.3 Two Drugs Rarely Given During Pregnancy – Predicted Harmful in Congenital Anomaly Model

Two drugs were rarely prescribed during pregnancy – Benzocaine mucosal spray and Hydromorphone Hydrochloride (Dilaudid), but were prescribed during the pre-conception period. This sudden drop-off in prescribing resulted in my algorithm detecting these two drugs as harmful (D or X) given that a similar drop-off in prescribing is often observed in known harmful (e.g., category X) drugs. Hydromorphone Hydrochloride (Dilaudid) is an opioid and therefore would typically not be prescribed during pregnancy given the harm that opioids are known to have on developing fetuses (Brennan and Rayburn, 2012). The other medication – benzocaine mucosal spray – is somewhat interesting. While a category C drug, benzocaine has been linked to development of methemoglobinemia in infants and because of the availability of safer category B medications is considered by many physicians to be contra-indicated during pregnancy according to several research papers (Lee et al., 2013; Peterson, 1960). My machine learning approach did not know this information *a priori*, but it was able to learn this from the physicians' usage pattern of the drug (i.e., dramatic drop-off of prescribing during pregnancy),

which is indicative of a category D or X drug. The ability to detect these types of potential D or X drugs is also of value to the pharmacology community since these drugs are treated as harmful by traditional standards of care.

7.5.2 Drugs Predicted Harmful in Fetal Loss Model

7.5.2.1 Drugs Treating Symptoms of Fetal Loss

Some drugs predicted to be harmful in the fetal loss cohort could have been prescribed to treat conditions leading up to a spontaneous abortion. For example, excessive bleeding often occurs during a spontaneous abortion, but this can take several days depending on the type of miscarriage. A drug used to treat severe bleeding following childbirth, or miscarriage, is Methylergonovine Maleate. All forms of Methylergonovine Maleate (3 different types listed in **Table 25**) had high rates of miscarriage following first trimester exposure – ranging from 97.9 – 100% of those exposed during that trimester. Typically, Methylergonovine Maleate would not be prescribed during the first trimester, unless something was wrong (e.g., excessive bleeding, which is indicative of a miscarriage). Therefore, this is likely a treatment-of-the-fetal-loss type of result.

7.5.2.2 Drugs That May Inadvertently Induce Fetal Loss: Diuretics

Furosemide is a potent diuretic with evidence of fetal-lethality among animals especially rabbits (FDA, 2012). Human studies involving fetal exposure to furosemide are sparse (i.e., indicative of a true FDA 'category C' drug). I found that 85.2% of pregnancies exposed during the first trimester resulted in fetal loss among the cohort. In general, diuretics are discouraged during pregnancy even as a treatment for hypertension when benefits may outweigh risks (Cunningham and Lindheimer, 1992), indicating the potential fetal harm of diuretic exposure, in general, on a developing fetus.

7.5.2.3 Drugs That May Inadvertently Induce Fetal Loss: DHCR7 Mechanism

Haloperidol was found to increase the risk of fetal loss following first trimester exposure. Haloperidol injection increased risk of fetal loss from 22.2% in the 3-6 months prior to conception to 78.6% if the exposure was in the first trimester. Nine pregnancies were exposed in the 3-6 month period pre-conception while 28 pregnancies were exposed during the first trimester – 22 resulted in fetal loss. Haloperidol is important as it is known to increase the expression of 7-dehydrocholesterol reductase (DHCR7) - an enzyme important in the conversion of 7-dehydrocholesterol to cholesterol (Boland and Tatonetti, 2016a). While exposure to pharmacological inhibitors of DHCR7 is known to result in fetal anomalies, the effects of drugs that merely increase the expression of the gene are less-well known. Drugs increasing expression of DHCR7 are not known to increase fetal loss (Boland and Tatonetti, 2016a). However, early fetal loss can be difficult to capture in many situations, whereas congenital anomalies in live-born babies is often easier to measure in case studies. DHCR7 is known to be important in regulating two very important processes for fetal development – cholesterol and vitamin D and perturbation of either of these important hormones is likely to adversely affect a fetus.

Some research suggests that neural toxicity of haloperidol can be rescued through a vitamin E mechanism (Behl et al., 1995). Variations in female's vitamin E concentrations could potentially explain variability observed in case studies and reported here with regards to fetal toxicity of haloperidol exposures. Therefore, increases in risk of fetal loss among first trimester exposure to haloperidol injections may be due to these underlying biochemical mechanisms.

7.5.3 Genetic Targets of Drugs More Predictive Than Classification

Prenatal vitamin supplementation is extremely important in reducing the overall disease risk of the offspring. For example, research has linked prenatal vitamin supplementation with lower

rates of leukemia, pediatric brain tumors and neuroblastoma (Goh et al., 2007). I restricted the analyses to identification of congenital anomalies diagnosed within the first 90 days of life. Therefore, I did not investigate complex outcomes such as childhood cancers or autism. However, vitamin-exposure during the prenatal period is widely considered to be an important variable in predicting fetal outcome. Hence, I included information on whether a drug was a vitamin or not in the models. In addition, I added information on whether a drug affected any vitamin-related protein. All models showed that knowing whether or not a drug affected a vitamin-related protein was more important then just knowing that a drug was a prenatal supplement (**Figure 33**). This is critical because it shows that a drug's mechanism of action and how it interfaces with vitamin-related mechanisms is extremely important in determine the fetal outcomes following drug exposure. This was known for certain drugs (Boland and Tatonetti, 2016a), but not across a larger cohort of fetal drug exposures. This has important clinical and pharmacological applications for future fetal toxicity studies.

7.6 Limitations

One drug predicted as harmful in the congenital anomalies model, was a cervical ripener - the drug implant Dinoprostone (or Cervidil). Cervical ripeners are drug-devices that are inserted into the vagina-cervix area for the induction of labor. It is possible that it was predicted as harmful because high-risk pregnancies are at increased risk of complications during delivery, which may lead to a higher-risk of congenital anomalies (Sunitha et al., 2017). My method identifies drugs predicted to be harmful given their prescribing patterns (e.g., low exposure during pregnancy), anomaly rates (e.g., proportion of exposed with an anomaly) and other factors often important in determining fetal effect (e.g., affecting proteins involved in vitamin-related processes). Further

study is needed to confirm that drugs predicted to be 'not-known-to-be-harmful' (i.e., similar to A or B category drugs) are in fact not harmful to the developing fetus.

7.7 Conclusion

In conclusion, I developed a machine learning approach that predicts drugs to be either harmful (D or X) or not-known-to-be-harmful (A or B) in two outcome models – fetal loss and congenital anomalies. Some drugs were predicted to be harmful because physicians stopped prescribing them upon pregnancy diagnosis – this dramatic drop-off in exposure rates triggered the algorithm to detect the drug as harmful (since a similar pattern is observed among drugs that are known to be harmful). Other drugs were predicted to be harmful because of the increase in anomalies observed following exposure - and many of these were validated in the literature, e.g., first trimester exposure to rubella live vaccine. Additionally, I found that first trimester exposure to haloperidol – a drug that interferes with the DHCR7 – cholesterol – vitamin D pathway increased the risk of fetal loss. The models achieved an OOB estimated accuracy of 90.6% for fetal loss and 87.1% for congenital anomalies. The model confirmed that NSAIDs – naproxen and ibuprofen – increased the risk of congenital anomalies, a finding reported in the literature. My approach provides much needed information for pharmacologists and prescribers interested in understanding the fetal effects of drugs.

7.8 Acknowledgments

This chapter is a reproduction, in whole or in part, of work to be submitted shortly (Boland et al., 2017d). Support for this research provided by the following sources. MRB was supported by the NCATS, NIH, through TL1 TR000082, formerly the NCRR, TL1 RR024158 from Jul. 2016 – Jun. 2017 when this work was conducted and by R01 GM107145. NPT was supported by: R01 GM107145, OT3 TR002027, and an award from the Herbert and Florence Irving Foundation.

Chapter 8

Concluding Remarks

In this dissertation, I have set out to explore the relationship between the seasonal factors of early development (both prenatal and perinatal) and how those seasonal factors affect later risk of disease. The goal of my work was to be applicable to both clinicians and pharmacologists by increasing the understanding of how these factors affect human health and disease.

While many birth month or birth season studies have been conducted throughout the world, no one previously has established whether or not birth season is causal in later risk of disease in a general sense. To that end, I have utilized nine criteria set forth by Dr. Hill to determine whether there is enough evidence underlying an association to suggest if it is causal or not (Hill, 1965). These nine criteria are known as Hill's criteria for causality. Various aims, described in separate chapters, address seven of nine of the Hill's criteria to establish that enough evidence does exist for a causal relationship between birth season and later risk of disease in a general sense.

The criteria addressed by this dissertation include: *strength, consistency, specificity, biological gradient, plausibility, coherence* and *experiment. Analogy* and *temporality* were not addressed. Testing for *analogy* was not necessary because phenocopies of the birth season – disease effect

can be directly tested using pharmacological inhibitors (typically *analogy* is used in lieu of other criteria). Additionally, I did not directly address *temporality*. Associations were tested using exposures taken at different sites to determine if the birth month – disease risk covaried along with the exposure. **Table 27** outlines the seven of nine Hill criteria addressed in this dissertation and how a link between birth month/seasonality and lifetime disease risk appears solid across many studies of diverse types conducted as a result of this dissertation.

Through this work, I identified several climate (e.g., rainfall, sunlight) and pollutant drivers (e.g., carbon monoxide) that were correlated with the birth season – disease effect across multiple sites throughout the world. These associations represent strong proof for the existence of birth season – disease relationships. In the second section of this dissertation, I explore mechanistic explanations for the population-level insights gained in the first half of the dissertation.

To that end, I first explored connections between genes involved in seasonal processes (e.g., vitamin D, folate regulation) and genes involved in diseases where risk is modulated by birth season (e.g., diabetes, hypertension). Then, I investigated the use of proxies (pharmacological inhibitors) to test retrospectively if the fetal outcomes mimicked what was observed in adults albeit more severe (Boland and Tatonetti, 2016a).

Drugs that target genes involved in vitamin D processes (specifically DHCR7) resulted in adverse fetal effects upon prenatal exposure. A deep exploration of drugs that inhibit DHCR7 showed increased fetal malformations and anomalies (Boland and Tatonetti, 2016a). Drugs that increased DHCR7 expression were shown to increase the risk of fetal loss ('miscarriage') following the development of a machine learning algorithm used on EHRs at CUMC (Boland et al., 2017d).

In addition, I was able to provide much needed information on the fetal toxicity or the lack

thereof for many drugs with no prior FDA recommendation with regards to their pregnancy status (i.e., FDA category C drugs). This information is very useful for clinicians, pharmacologists and women considering taking a medication during pregnancy that is determined to be category C. This included category C drugs that did not necessarily target a known seasonal gene/protein. This is also important because the pharmacological action of drugs is still in the nascent stage. Many of these drugs identified to be harmful through my machine learning approach may in fact target vitamin – related processes that have yet to be identified at this early stage. This dissertation establishes that birth month / season is an important factor in later risk of disease. Further, pharmacological perturbation of proteins undergoing seasonal regulation increases the risk of fetal harm. Informatics-methods can be used to combine both population-level analyses and mechanistic approaches to understand important physiological processes.

Life is short, and Art long; the crisis fleeting; experience perilous, and decision difficult.

~Hippocrates

Table 27. This Dissertation In the Context of Hill's Nine Criteria:

Determining Causality for the Birth Month – Health Relationship

Criterion	Desc.	Desc. As Related to Birth Season Effect	Aim	Ch.	Reference
Strength	How strong is the association?	What is the Relative Risk by birth month for disease associations?	1	2	(Boland et al., 2015b)
Consistency	Is the association repeatedly observed by different persons, places, and times	Are findings replicated at other sites?	1	2,3	(Li et al., 2016) (Boland et al., 2017b)
Specificity	Is the association between occupation x and disease y specific? Or is a plethora of diseases associated with occupation x?	Do disease categories cluster around certain birth seasons?	1	2	(Boland et al., 2015b)
Temporality	Does the cart come before the horse? Or the horse before the cart? What items are the cart and what is the horse	Does sunlight exposure (or other climate variable) precede disease risk curve?	-		-
Biological gradient	Dose-response	Does increasing the particular climate variable (e.g., sunlight) correspondingly affect the birth season – disease effect?	2	3	(Boland et al., 2017b)
Plausibility	Biological plausibility – is there a known mechanism that can explain the observation?	Does the effect we observe fit with what is known about the particular climate variable (e.g., sunlight) and its effect on the body and prenatal gene expression	3	5	(Boland and Tatonetti, 2016c)
Coherence	All findings related to the association should contribute to the overall mechanism involved (adds weight to the mechanism being correct)	When associations are found or failed to be found at certain sites – does it fit with what we know about the climate drivers underlying the relationship?	2	3	(Boland et al., 2017c); (Boland et al., 2015a)
Experiment	Direct experimental evidence (gene knockout, phenocopy of birth season effect)	Using drugs as a phenocopy for the birth season – disease effect, can we mimic the climate driver's effect on prenatal outcomes through the use of a pharmacological inhibitor?	4	6, 7	(Boland and Tatonetti, 2016a) (Boland et al., 2017d)
Analogy	We can learn from thalidomide and rubella and through the use of analogy accept slighter but similar evidence with another drug or viral disease when exposed during pregnancy.	Perhaps compare birth season effect to drug exposures during pregnancy through use of the phenocopy approach – however this can be done directly (see *Experiment*) and use of *Analogy* is not necessary	-		-

Bibliography

(2014a) Average sunshine in New York, United States of America. *World Weather and Climate Information* http://www.weather-and-climate.com/average-monthly-hours-Sunshine,New-York,United-States-of-America.

(2014b) Average sunshine in Skagen, Denmark. *World Weather and Climate Information* http://www.weather-and-climate.com/average-monthly-hours-Sunshine,skagen,Denmark.

Aarskog D (1979) Maternal Progestins as a Possible Cause of Hypospadias. *New England Journal of Medicine* **300**:75-78.

Åberg N (1989) Birth season variation in asthma and allergic rhinitis. *Clinical & Experimental Allergy* **19**:643-648.

Ahn H, Choi J, Han J, Kim M, Chung J, Ryu H, Kim M, Yang J, Koong M and Nava-Ocampo A (2008) Pregnancy outcome after exposure to oral contraceptives during the periconceptional period. *Human & experimental toxicology* **27**:307-313.

Ahn J, Yu K, Stolzenberg-Solomon R, Simon KC, McCullough ML, Gallicchio L, Jacobs EJ, Ascherio A, Helzlsouer K, Jacobs KB, Li Q, Weinstein SJ, Purdue M, Virtamo J, Horst R, Wheeler W, Chanock S, Hunter DJ, Hayes RB, Kraft P and Albanes D (2010) Genome-wide association study of circulating vitamin D levels. *Human Molecular Genetics* **19**:2739-2745.

Al-Owain M, Imtiaz F, Shuaib T, Edrees A, Al-Amoudi M, Sakati N, Al-Hassnan Z, Bamashmous H, Rahbeeni Z and Al-Ameer S (2012) Smith–Lemli–Opitz syndrome among Arabs. *Clinical genetics* **82**:165-172.

Allen JP (2005) *The art of medicine in ancient Egypt*, Metropolitan Museum of Art.

Anai T, Matsu T, Oga M, Yoshimatsu J and Miyakawa I (1991) Seasonal incidence of subclinical vitamin K deficiency during early newborn period. *Nihon Sanka Fujinka Gakkai zasshi* **43**:342-346.

Andrade SE, Gurwitz JH, Davis RL, Chan KA, Finkelstein JA, Fortman K, McPhillips H, Raebel MA, Roblin D and Smith DH (2004) Prescription drug use in pregnancy. *American journal of obstetrics and gynecology* **191**:398-407.

Andreadis C, Charalampidou M, Diamantopoulos N, Chouchos N and Mouratidou D (2004) Combined chemotherapy and radiotherapy during conception and first two trimesters of gestation in a woman with metastatic breast cancer. *Gynecologic oncology* **95**:252-255.

Anstey AV, Azurdia RM, Rhodes LE, Pearse AD and Bowden PE (2005) Photosensitive Smith–Lemli–Opitz syndrome is not caused by a single gene mutation: analysis of the gene

encoding 7-dehydrocholesterol reductase in five U.K. families. *British Journal of Dermatology* **153**:774-779.
Baker T (1963) A quantitative and cytological study of germ cells in human ovaries. *Proceedings of the royal society of london Series b, biological sciences*:417-433.
Bánhidy F, Puhó E and Czeizel AE (2005) Maternal influenza during pregnancy and risk of congenital abnormalities in offspring. *Birth Defects Research Part A: Clinical and Molecular Teratology* **73**:989-996.
Barni S, Ardizzoia A, Zanetta G, Strocchi E, Lissoni P and Tancini G (1992) Weekly doxorubicin chemotherapy for breast cancer in pregnancy. A case report. *Tumori* **78**:349-350.
Barreca AI and Shimshack JP (2012) Absolute Humidity, Temperature, and Influenza Mortality: 30 Years of County-Level Evidence from the United States. *American Journal of Epidemiology* **176**:S114-S122.
Barthelmes L and Gateley C (2004) Tamoxifen and pregnancy. *The Breast* **13**:446-451.
Basu T, Donald E, Hargreaves J, Thompson G, Chao E and Peterson R (1994) Seasonal variation of vitamin A (retinol) status in older men and women. *Journal of the American College of Nutrition* **13**:641-645.
Bates CJ, Prentice AM and Paul A (1994) Seasonal variations in vitamins A, C, riboflavin and folate intakes and status of pregnant and lactating women in a rural Gambian community: some possible implications. *European journal of clinical nutrition* **48**:660-668.
Becher H, Müller O, Jahn A, Gbangou A, Kynast-Wolf G and Kouyaté B (2004) Risk factors of infant and child mortality in rural Burkina Faso. *Bulletin of the World Health Organization* **82**:265-273.
Becker A, Eyles DW, McGrath JJ and Grecksch G (2005) Transient prenatal vitamin D deficiency is associated with subtle alterations in learning and memory functions in adult rats. *Behavioural Brain Research* **161**:306-312.
Beckman D and Brent R (1984) Mechanisms of teratogenesis. *Annual review of pharmacology and toxicology* **24**:483-500.
Behl C, Rupprecht R, Skutella T and Holsboer F (1995) Haloperidol-induced cell death-mechanism and protection with vitamin E in vitro. *Neuroreport* **7**:360-364.
Ben-Amotz A, Yatziv S, Sela M, Greenberg S, Rachmilevich B, Shwarzman M and Weshler Ze (1998) Effect of natural b-carotene supplementation in children exposed to radiation from the Chernobyl accident. *Radiation and Environmental Biophysics* **37**:187-193.
Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB and Levy D (1998) Impact of Atrial Fibrillation on the Risk of Death: The Framingham Heart Study. *Circulation* **98**:946-952.
Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*:289-300.
Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L and McGlinchey S (2014) The ChEMBL bioactivity database: an update. *Nucleic acids research* **42**:D1083-D1090.
Bérard A, Ramos E, Rey E, Blais L, St-André M and Oraichi D (2007) First trimester exposure to paroxetine and risk of cardiac malformations in infants: the importance of dosage. *Birth Defects Research Part B: Developmental and Reproductive Toxicology* **80**:18-27.

Berger JC and Clericuzio CL (2008) Pierre Robin sequence associated with first trimester fetal tamoxifen exposure. *American Journal of Medical Genetics Part A* **146**:2141-2144.

Bignon E, Pons M, Crastes de Paulet AC, Dore JC, Gilbert J, Abecassis J, Miquel JF, Ojasoo T and Raynaud JP (1989) Effect of triphenylacrylonitrile derivatives on estradiol-receptor binding and on human breast cancer cell growth. *Journal of medicinal chemistry* **32**:2092-2103.

Bjørnerem As, Straume B, Øian Pl and Berntsen GK (2006) Seasonal variation of estradiol, follicle stimulating hormone, and dehydroepiandrosterone sulfate in women and men. *The Journal of Clinical Endocrinology & Metabolism* **91**:3798-3802.

Bjornsdottir US, Holgate ST, Reddy PS, Hill AA, McKee CM, Csimma CI, Weaver AA, Legault HM, Small CG and Ramsey RC (2011) Pathways activated during human asthma exacerbation as revealed by gene expression patterns in blood. *PloS one* **6**:e21902.

Blair David R, Lyttle Christopher S, Mortensen Jonathan M, Bearden Charles F, Jensen Anders B, Khiabanian H, Melamed R, Rabadan R, Bernstam Elmer V, Brunak S, Jensen Lars J, Nicolae D, Shah Nigam H, Grossman Robert L, Cox Nancy J, White Kevin P and Rzhetsky A (2013) A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* **155**:70-80.

Blair SN, Kampert JB, Kohl HW, Iii and et al. (1996) Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *JAMA* **276**:205-210.

Boekelheide K, Blumberg B, Chapin RE, Cote I, Graziano JH, Janesick A, Lane R, Lillycrop K, Myatt L and Thayer KA (2012) Predicting later-life outcomes of early-life exposures. *Environmental health perspectives* **120**:1353.

Boland MR (2015) SeaWAS project code. *GitHub repository* <https://github.com/maryreginaboland/SeaWAS>

Boland MR, Hripcsak G, Albers DJ, Wei Y, Wilcox AB, Wei J, Li J, Lin S, Breene M, Myers R, Zimmerman J, Papapanou PN and Weng C (2013a) Discovering medical conditions associated with periodontitis using linked electronic health records. *J Clin Periodontol* **40**:474-482.

Boland MR, Hripcsak G, Ryan P and Tatonetti NP (2015a) A Climate-Wide Journey to Explore Mechanisms Underlying Birth Month-Disease Risk Associations: A Call for Collaboration. *Observational Health Data Sciences and Informatics Symposium* **Washington DC**.

Boland MR, Hripcsak G, Shen Y, Chung WK and Weng C (2013b) Defining a comprehensive verotype using electronic health records for personalized medicine. *J Am Med Inform Assoc* **20**:e232-238.

Boland MR, Karczewski KJ and Tatonetti NP (2017a) Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. *PLoS Comput Biol* **13**:e1005278.

Boland MR, Parhi P, Gentine P and Tatonetti NP (2017b) Climate Classification is an Important Factor in Assessing Quality-of-Care Across Hospitals. *In Press*.

Boland MR, Parhi P, Li L, Miotto R, Carroll R, Iqbal U, Nguyen A, Schuemie M, You SC, Ryan P, Li J, Park RW, Denny JC, Dudley JT, Hripcsak G, Gentine P and Tatonetti NP (2017c) Uncovering Exposures Responsible for Birth Season – Disease Effects: A Global Study. *Submitted*.

Boland MR, Polubriaginof F and Tatonetti NP (2017d) Development of A Machine Learning Algorithm to Classify Drugs Of Unknown Fetal Effect. *Submitted*.

Boland MR, Shahn Z, Madigan D, Hripcsak G and Tatonetti NP (2015b) Birth month affects lifetime disease risk: a phenome-wide method. *Journal of the American Medical Informatics Association* **22**:1042-1053.

Boland MR and Tatonetti N (2016a) Investigation of 7-dehydrocholesterol reductase pathway to elucidate off-target prenatal effects of pharmaceuticals: a systematic review. *The pharmacogenomics journal* **16**:411–429.

Boland MR and Tatonetti NP (2015) Are All Vaccines Created Equal? Using Electronic Health Records to Discover Vaccines Associated With Clinician-Coded Adverse Events. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science* **2015**:196-200.

Boland MR and Tatonetti NP (2016b) Assessing the Mutational Spectrum of 7-DeHydroCholesterol Reductase and the Toxicological Effects of Pharmacological Inhibition During the Prenatal Period. *Abstract / Prenatal, Perinatal and Reproductive Genetics #3260T* **Presented at the 66th Annual Meeting of The American Society of Human Genetics, October 20, 2016, Vancouver, Canada.**

Boland MR and Tatonetti NP (2016c) In Search of 'Birth Month Genes': Using Existing Data Repositories to Locate Genes Underlying Birth Month-Disease Relationships. *AMIA Summits on Translational Science Proceedings* **189-198**.

Boothby LA and Doering PL (2001) FDA labeling system for drugs in pregnancy. *Annals of Pharmacotherapy* **35**:1485-1489.

Boulding W, Glickman SW, Manary MP, Schulman KA and Staelin R (2011) Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. *The American journal of managed care* **17**:41-48.

Braems G, Denys H, De Wever O, Cocquyt V and Van den Broecke R (2011) Use of tamoxifen before and during pregnancy. *The oncologist* **16**:1547-1551.

Brennan MC and Rayburn WF (2012) Counseling about risks of congenital anomalies from prescription opioids. *Birth Defects Research Part A: Clinical and Molecular Teratology* **94**:620-625.

Brook RD, Bard RL, Burnett RT, Shin HH, Vette A, Croghan C, Phillips M, Rodes C, Thornburg J and Williams R (2010) Differences in blood pressure and vascular responses associated with ambient fine particulate matter exposures measured at the personal versus community level. *Occupational and environmental medicine*:oem. 2009.053991.

Bryant HE, Visser N and Love EJ (1989) Records, recall loss, and recall bias in pregnancy: a comparison of interview and medical records data of pregnant and postnatal women. *American Journal of Public Health* **79**:78-80.

Bunnage ME, Gilbert AM, Jones LH and Hett EC (2015) Know your target, know your molecule. *Nat Chem Biol* **11**:368-372.

Burga A and Lehner B (2012) Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *FEBS Journal* **279**:3765-3775.

Burke PJ, Kalet BT and Koch TH (2004) Antiestrogen binding site and estrogen receptor mediate uptake and distribution of 4-hydroxytamoxifen-targeted doxorubicin-formaldehyde conjugate in breast cancer cells. *Journal of medicinal chemistry* **47**:6509-6518.

Burne THJ, Féron F, Brown J, Eyles DW, McGrath JJ and Mackay-Sim A (2004) Combined prenatal and chronic postnatal vitamin D deficiency in rats impairs prepulse inhibition of acoustic startle. *Physiology & Behavior* **81**:651-655.
Busse W and Sedgwick J (1992) Eosinophils in asthma. *Annals of allergy* **68**:286-290.
Calderón J, Navarro ME, Jimenez-Capdeville ME, Santos-Diaz MA, Golden A, Rodriguez-Leyva I, Borja-Aburto V and Díaz-Barriga F (2001) Exposure to Arsenic and Lead and Neuropsychological Development in Mexican Children. *Environmental Research* **85**:69-76.
Campbell JR and Payne TH (1994) A comparison of four schemes for codification of problem lists. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care*:201-205.
Cantorna MT, Zhu Y, Froicu M and Wittke A (2004) Vitamin D status, 1,25-dihydroxyvitamin D3, and the immune system. *The American Journal of Clinical Nutrition* **80**:1717S-1720S.
Cardonick E, Gilmandyar D and Somer RA (2012) Maternal and neonatal outcomes of dose-dense chemotherapy for breast cancer in pregnancy. *Obstetrics & Gynecology* **120**:1267-1272.
Cardoso M, Balreira A, Martins E, Nunes L, Cabral A, Marques M, Lima MR, Marques J, Medeira A and Cordeiro I (2005) Molecular studies in Portuguese patients with Smith–Lemli–Opitz syndrome and report of three new mutations in DHCR7. *Molecular genetics and metabolism* **85**:228-235.
Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P and Riddell A (2016) Stan: A probabilistic programming language. *J Stat Softw*.
Carr D, Whiteley G, Alfirevic A and Pirmohamed M (2009) Investigation of inter-individual variability of the one-carbon folate pathway: a bioinformatic and genetic review. *The pharmacogenomics journal* **9**:291-305.
CDC (2008) Update on overall prevalence of major birth defects--Atlanta, Georgia, 1978-2005. *MMWR Morbidity and mortality weekly report* **57**:1.
CDC (2014a) The Flu Season. http://www.cdc.gov/flu/about/season/flu-season.htm.
CDC (2014b) Vital Stats Beyond 20/20. *National Vital Statistics System US Department of Health and Human Services* http://205.207.175.93/Vitalstats/Common/Login/Login.aspx.
CDC (2015) Measuring Gestational Age in Vital Statistics Data: Transitioning to the Obstetric Estimate. *National Vital Statistics Reports* <http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_05.pdf>
CDC (2016) Most Recent Asthma Data: National and State Data. *Accessed on May 24, 2016* <http://www.cdc.gov/asthma/most_recent_data.htm>.
CDC (2017) CDCs Abortion Surveillance System FAQs: Abortion Surveillance—Findings and Reports. *Reproductive Health* <https://www.cdc.gov/reproductivehealth/data_stats/abortion.htm>
CDC, Ventura SJ, Curtin SC, Abma JC and Henshaw SK (2012) Estimated Pregnancy Rates and Rates of Pregnancy Outcomes for the United States, 1990–2008. *National Vital Statistics Reports* <https://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_07.pdf>.
Chang LC, Bhat KP, Pisha E, Kennelly EJ, Fong HH, Pezzuto JM and Kinghorn AD (1998) Activity-guided isolation of steroidal alkaloid antiestrogen-binding site inhibitors from Pachysandra procumbens. *Journal of natural products* **61**:1257-1262.

Chatterjee A and Mukheree J (1997) Comparative study of different anticonvulsants in eclampsia. *Journal of Obstetrics and Gynaecology Research* **23**:289-293.
Chen C-J, Chiou H-Y, Chiang M-H, Lin L-J and Tai T-Y (1996) Dose-Response Relationship Between Ischemic Heart Disease Mortality and Long-term Arsenic Exposure. *Arteriosclerosis, Thrombosis, and Vascular Biology* **16**:504-510.
Chen J, Radford MJ, Wang Y, Marciniak TA and Krumholz HM (1999) Do "America's Best Hospitals" Perform Better for Acute Myocardial Infarction? *New England Journal of Medicine* **340**:286-292.
Chen M-H, Lan W-H, Bai Y-M, Huang K-L, Su T-P, Tsai S-J, Li C-T, Lin W-C, Chang W-H and Pan T-L (2016) Influence of relative age on diagnosis and treatment of attention-deficit hyperactivity disorder in Taiwanese children. *The Journal of pediatrics* **172**:162-167. e161.
Chen M-H, Su T-P, Chen Y-S, Hsu J-W, Huang K-L, Chang W-H, Chen T-J and Bai Y-M (2013) Asthma and attention-deficit/hyperactivity disorder: a nationwide population-based prospective cohort study. *Journal of Child Psychology and Psychiatry* **54**:1208-1214.
Chin DL, Bang H, Manickam RN and Romano PS (2016) Rethinking Thirty-Day Hospital Readmissions: Shorter Intervals Might Be Better Indicators Of Quality Of Care. *Health Affairs* **35**:1867-1875.
Chiou H-Y, Hsueh Y-M, Liaw K-F, Horng S-F, Chiang M-H, Pu Y-S, Shinn-Nan Lin J, Huang C-H and Chen C-J (1995) Incidence of Internal Cancers and Ingested Inorganic Arsenic: A Seven-Year Follow-up Study in Taiwan. *Cancer Research* **55**:1296-1300.
Clapp R and Ozonoff D (2000) Where the boys aren't: dioxin and the sex ratio. *The Lancet* **355**:1838-1839.
Clark S (1993) Prophylactic tamoxifen. *The Lancet* **342**:168.
Clausen TD, Mathiesen ER, Hansen T, Pedersen O, Jensen DM, Lauenborg J and Damm P (2008) High prevalence of type 2 diabetes and pre-diabetes in adult offspring of women with gestational diabetes mellitus or type 1 diabetes the role of intrauterine hyperglycemia. *Diabetes care* **31**:340-346.
CMS (2015) Hospital Compare data archive. *DataMedicareGov* https://data.medicare.gov/data/archives/hospital-compare
Cohen HA, Blau H, Hoshen M, Batat E and Balicer RD (2014) Seasonality of asthma: a retrospective population study. *Pediatrics* **133**:e923-e932.
Cooper LZ and Krugman S (1967) Clinical manifestations of postnatal and congenital rubella. *Archives of Ophthalmology* **77**:434-439.
Cooper RS (2001) Social inequality, ethnicity and cardiovascular disease. *International Journal of Epidemiology* **30**:S48.
Correa-Cerro LS, Wassif CA, Kratz L, Miller GF, Munasinghe JP, Grinberg A, Fliesler SJ and Porter FD (2006) Development and characterization of a hypomorphic Smith–Lemli–Opitz syndrome mouse model and efficacy of simvastatin therapy. *Human Molecular Genetics* **15**:839-851.
Correa-Cerro LS, Wassif CA, Waye JS, Krakowiak PA, Cozma D, Dobson NR, Levin SW, Anadiotis G, Steiner RD, Krajewska-Walasek M, Nowaczyk MJM and Porter FD (2005) DHCR7 nonsense mutations and characterisation of mRNA nonsense mediated decay in Smith-Lemli-Opitz syndrome. *Journal of Medical Genetics* **42**:350-357.

Coventry PA, Gemmell I and Todd CJ (2011) Psychosocial risk factors for hospital readmission in COPD patients on early discharge services: a cohort study. *BMC Pulmonary Medicine* **11**:1-10.
Crawford DC, Crosslin DR, Tromp G, Kullo IJ, Kuivaniemi H, Hayes MG, Denny JC, Bush WS, Haines JL and Roden DM (2014) eMERGEing progress in genomics—the first seven years. *Frontiers in genetics* **5**.
Cross J, Iben J, Simpson C, Thurm A, Swedo S, Tierney E, Bailey-Wilson J, Biesecker L, Porter F and Wassif C (2014) Determination of the allelic frequency in Smith–Lemli–Opitz syndrome by analysis of massively parallel sequencing data sets. *Clinical genetics*.
Crowther CA (1985) Eclampsia at Harare Maternity Hospital: an epidemiological study. *South African Medical Journal* **68**:927-929.
CTD (2015) Comparative Toxicogenomics Database. http://ctdbase.org/. **Accessed in February 2015**.
Cullins SL, Pridjian G and Sutherland CM (1994) Goldenhar's syndrome associated with tamoxifen given to the mother during gestation. *Jama* **271**:1905-1906.
Cunningham FG and Lindheimer MD (1992) Hypertension in pregnancy. *New England Journal of Medicine* **326**:927-932.
Czeizel AE and Dudas I (1992) Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *New England journal of medicine* **327**:1832-1835.
Dai WS, LaBraico JM and Stern RS (1992) Epidemiology of isotretinoin exposure during pregnancy. *Journal of the American Academy of Dermatology* **26**:599-606.
Dally A (1998) Thalidomide: was the tragedy preventable? *The Lancet* **351**:1197.
Das J and Mohpal A (2016) Socioeconomic Status And Quality Of Care In Rural India: New Evidence From Provider And Household Surveys. *Health Affairs* **35**:1764-1773.
Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Wiegers TC and Mattingly CJ (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Research* **43**:D914-D920.
Davis RE, Rossier CE and Enfield KB (2012) The Impact of Weather on Influenza and Pneumonia Mortality in New York City, 1975?2002: A Retrospective Study. *PLoS ONE* **7**:e34091.
Dawson-Hughes B, Dallal GE, Krall EA, Harris S, Sokoll LJ and Falconer G (1991) Effect of Vitamin D Supplementation on Wintertime and Overall Bone Loss in Healthy Postmenopausal Women. *Annals of Internal Medicine* **115**:505-512.
De Brasi D, Esposito T, Rossi M, Parenti G, Sperandeo M, Zuppaldi A, Bardaro T, Ambruzzi M, Zelante L and Ciccodicola A (1999) Smith-Lemli-Opitz syndrome: evidence of T93M as a common mutation of delta7-sterol reductase in Italy and report of three novel mutations. *European journal of human genetics : EJHG* **7**:937-940.
de Wildt SN, Taguchi N and Koren G (2009) Unintended pregnancy during radiotherapy for cancer. *Nature clinical practice Oncology* **6**:175-178.
De-Regil LM, Palacios C, Ansary A, Kulier R and Pena-Rosas JP (2012) Vitamin D supplementation for women during pregnancy. *Cochrane Database Syst Rev* **2**.
De-Regil LM, Palacios C, Lombardo LK and Peña-Rosas JP (2016) Vitamin D supplementation for women during pregnancy. *Cochrane Database of Systematic Reviews*.
Deeb KK, Trump DL and Johnson CS (2007) Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat Rev Cancer* **7**:684-700.

Dell M, Jones BF and Olken BA (2012) Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*:66-95.
Dell M, Jones BF and Olken BA (2013) What do we learn from the weather? The new climate-economy literature, National Bureau of Economic Research.
Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM and Crawford DC (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**:1205-1210.
DerSimonian R and Laird N (1986) Meta-analysis in clinical trials. *Controlled clinical trials* **7**:177-188.
Diav-Citrin O, Shechtman S, Ornoy S, Arnon J, Schaefer C, Garbis H, Clementi M and Ornoy A (2005) Safety of haloperidol and penfluridol in pregnancy: a multicenter, prospective, controlled study. *The Journal of clinical psychiatry* **66**:317-322.
Dickersin K (1990) The existence of publication bias and risk factors for its occurrence. *JAMA* **263**:1385-1389.
Disanto G, Chaplin G, Morahan JM, Giovannoni G, Hypponen E, Ebers GC and Ramagopalan SV (2012) Month of birth, vitamin D and risk of immune mediated disease: a case control study. *BMC medicine* **10**:69.
Doblhammer G and Vaupel JW (2001) Lifespan depends on month of birth. *Proceedings of the National Academy of Sciences* **98**:2934-2939.
Dominici F, Peng RD, Bell ML and et al. (2006) FIne particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **295**:1127-1134.
Dopico XC, Evangelou M, Ferreira RC, Guo H, Pekalski ML, Smyth DJ, Cooper N, Burren OS, Fulford AJ and Hennig BJ (2015) Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nature communications* **6**.
Doshi-Velez F, Ge Y and Kohane I (2014) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* **133**:e54-63.
Douglas AS (1993) Seasonality of Hip Fracture and Haemorrhagic Disease of the Newborn. *Scottish Medical Journal* **38**:37-40.
Duffy MR, Chen T-H, Hancock WT, Powers AM, Kool JL, Lanciotti RS, Pretrick M, Marfel M, Holzbauer S and Dubray C (2009) Zika virus outbreak on Yap Island, federated states of Micronesia. *New England Journal of Medicine* **360**:2536-2543.
Duncan J, Narus SP, Clyde S, Eilbeck K, Thornton S and Staes C (2014) Birth of identity: understanding changes to birth certificates and their value for identity resolution. *J Am Med Inform Assoc*.
Easterbrook PJ, Gopalan R, Berlin JA and Matthews DR (1991) Publication bias in clinical research. *The Lancet* **337**:867-872.
Edison RJ and Muenke M (2004a) Central Nervous System and Limb Anomalies in Case Reports of First-Trimester Statin Exposure. *New England Journal of Medicine* **350**:1579-1582.
Edison RJ and Muenke M (2004b) Mechanistic and epidemiologic considerations in the evaluation of adverse birth outcomes following gestational exposure to statins. *American Journal of Medical Genetics Part A* **131**:287-298.
Edison RJ and Muenke M (2005) Gestational Exposure to Lovastatin Followed by Cardiac Malformation Misclassified as Holoprosencephaly. *New England Journal of Medicine* **352**:2759-2759.

Egger G, Liang G, Aparicio A and Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**:457-463.

Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST and Speroff T (2006) Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. *Mayo Clinic Proceedings* **81**:741-748.

Engelsman E, Heuson JC, Blonk Van Der Wijst J, Drochmans A, Maass H, Cheix F, Sobrinho LG and Nowakowski H (1975) Controlled clinical trial of L-dopa and nafoxidine in advanced breast cancer: an E.O.R.T.C. study. *British medical journal* **2**:714-715.

Epstein PR (1999) Climate and health. *Science* **285**:347.

Ergenoglu AM, Yeniel AÖ, Yildirim N, Kazandi M, Akercan F and Sagol S (2012) Rubella vaccination during the preconception period or in pregnancy and perinatal and fetal outcomes. *The Turkish journal of pediatrics* **54**:230.

Ericson A and Källén BAJ (2001) Nonsteroidal anti-inflammatory drugs in early pregnancy. *Reproductive Toxicology* **15**:371-375.

Etterson JR, Schneider HE, Gorden NLS and Weber JJ (2016) Evolutionary insights from studies of geographic variation: Contemporary variation and looking to the future. *American Journal of Botany* **103**:5-9.

Evans T, Poh A, Webb C, Wainwright B, Wicking C, Glass I, Carey WF and Fietz M (2001) Novel mutation in the Δ7-dehydrocholesterol reductase gene in an Australian patient with Smith-Lemli-Opitz syndrome. *American Journal of Medical Genetics* **103**:344-347.

Evans WN, Morrill MS and Parente ST (2010) Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children. *Journal of health economics* **29**:657-673.

FDA (2012) LASIX (furosemide) Tables 20, 40 and 80 mg Warning Label Information. <http://wwwaccessdatafdagov/drugsatfda_docs/label/2012/016273s066lblpdf> **Accessed on March 2, 2017**.

Fernø J, Raeder M, Vik-Mo A, Skrede S, Glambek M, Tronstad K, Breilid H, Løvlie R, Berge R and Stansberg C (2005) Antipsychotic drugs activate SREBP-regulated expression of lipid biosynthetic genes in cultured human glioma cells: a novel mechanism of action? *The pharmacogenomics journal* **5**:298-304.

Ferno J, Skrede S, Vik-Mo A, Havik B and Steen V (2006) Drug-induced activation of SREBP-controlled lipogenic gene expression in CNS-related cell lines: Marked differences between various antipsychotic drugs. *BMC Neuroscience* **7**:69.

Finnell RH (1999) Teratology: General considerations and principles. *Journal of Allergy and Clinical Immunology* **103**:S337-S342.

Fishman MC and Porter JA (2005) Pharmaceuticals: A new grammar for drug discovery. *Nature* **437**:491-493.

Fitzky BU, Witsch-Baumgartner M, Erdel M, Lee JN, Paik Y-K, Glossmann H, Utermann G and Moebius FF (1998) Mutations in the Δ7-sterol reductase gene in patients with the Smith–Lemli–Opitz syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **95**:8181-8186.

Foley D and Mackinnon A (2014) A systematic review of antipsychotic drug effects on human gene expression related to risk factors for cardiovascular disease. *The pharmacogenomics journal* **14**:446-451.

Fricke-Galindo I, Cespedes-Garro C, Rodrigues-Soares F, Naranjo MEG, Delgado A, de Andres F, Lopez-Lopez M, Penas-Lledo E and Llerena A (2016) Interethnic variation of CYP2C19 alleles, /`predicted/' phenotypes and /`measured/' metabolic phenotypes across world populations. *Pharmacogenomics J* **16**:113-123.

Frickhofen N, Abkowitz JL, Safford M, Berry JM, Antunez-de-Mayolo J, Astrow A, Cohen R, Halperin I, King L, Mintzer D, Cohen B and Young NS (1990) Persistent B19 Parvovirus Infection in Patients Infected with Human Immunodeficiency Virus Type 1 (HIV-1): A Treatable Cause of Anemia in AIDS. *Annals of Internal Medicine* **113**:926-933.

Fujimori K, Kyozuka H, Yasuda S, Goto A, Yasumura S, Ota M, Ohtsuru A, Nomura Y, Hata K and Suzuki K (2014) Pregnancy and birth survey after the great East Japan earthquake and fukushima daiichi nuclear power plant accident in fukushima prefecture. *Fukushima journal of medical science* **60**:75-81.

Fukazawa R, Nakahori Y, Kogo T, Kawakami T, Akamatsu H, Tanae A, Hibi I, Nagafuchi S, Nakagome Y and Hirayama T (1992) Normal Y sequences in Smith-Lemli-Opitz syndrome with total failure of masculinization. *Acta paediatrica* **81**:570-572.

Gelardi M, Peroni DG, Incorvaia C, Quaranta N, De Luca C, Barberi S, Dell'Albani I, Landi M, Frati F and de Beaumont O (2014) Seasonal changes in nasal cytology in mite-allergic patients. *Journal of inflammation research* **7**:39.

Gelman A and Price PN (1999) All maps of parameter estimates are misleading. *Statistics in medicine* **18**:3221-3234.

Gentile S (2004) Clinical utilization of atypical antipsychotics in pregnancy and lactation. *Annals of Pharmacotherapy* **38**:1265-1271.

Germann N, Goffinet F and Goldwasser F (2004) Anthracyclines during pregnancy: embryo–fetal outcome in 160 patients. *Annals of Oncology* **15**:146-150.

Ginat S, Battaile KP, Battaile BC, Maslen C, Gibson KM and Steiner RD (2004) Lowered DHCR7 activity measured by ergosterol conversion in multiple cell types in Smith–Lemli–Opitz syndrome. *Molecular Genetics and Metabolism* **83**:175-183.

Glance LG, Kellermann AL, Osler TM, Li Y, Li W and Dick AW (2015) Impact of Risk Adjustment for Socioeconomic Status on Risk-adjusted Surgical Readmission Rates. *Annals of surgery*.

Goh Y, Bollano E, Einarson T and Koren G (2007) Prenatal multivitamin supplementation and rates of pediatric cancers: a meta-analysis. *Clinical Pharmacology & Therapeutics* **81**:685-691.

Goldenberg A, Chevy F, Bernard C, Wolf C and Cormier-Daire V (2003) Circonstances cliniques du diagnostic du syndrome de Smith-Lemli-Opitz et tentatives de corrélation phénotype-génotype : à propos de 45 cas. *Archives de Pédiatrie* **10**:4-10.

Goldstein DJ (1995) Effects of third trimester fluoxetine exposure on the newborn. *Journal of clinical psychopharmacology* **15**:417-420.

Green EL (1968) Genetic effects of radiation on mammalian populations. *Annual Review of Genetics* **2**:87-120.

Group EBC (1972) Clinical trial of nafoxidine, an oestrogen antagonist in advanced breast cancer. *European Journal of Cancer (1965)* **8**:387-389.

Grzybowska E, Hemminki K, Szeliga J and Chorazy M (1993) Seasonal variation of aromatic DNA adducts in human lymphocytes and granulocytes. *Carcinogenesis* **14**:2523-2526.

Guo L, Xiao Y and Wang Y (2013) Hexavalent Chromium-induced Alteration of Proteomic Landscape in Human Skin Fibroblast Cells. *Journal of Proteome Research* **12**:3511-3518.

Guo L, Xiao Y and Wang Y (2014) Monomethylarsonous acid inhibited endogenous cholesterol biosynthesis in human skin fibroblasts. *Toxicology and Applied Pharmacology* **277**:21-29.

Haerian K, Varn D, Vaidya S, Ena L, Chase H and Friedman C (2012) Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology & Therapeutics* **92**:228-234.

Hahn KM, Johnson PH, Gordon N, Kuerer H, Middleton L, Ramirez M, Yang W, Perkins G, Hortobagyi GN and Theriault RL (2006) Treatment of pregnant breast cancer patients and outcomes of children exposed to chemotherapy in utero. *Cancer* **107**:1219-1226.

Halasyamani LK and Davis MM (2007) Conflicting measures of hospital quality: ratings from "Hospital Compare" versus "Best Hospitals". *Journal of Hospital Medicine* **2**:128-134.

Hales C and Barker D (2013) Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *International journal of epidemiology* **42**:1215-1222.

Hales CN and Barker DJ (1992) Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* **35**:595-601.

Halicioglu O, Sutcuoglu S, Koc F, Yildiz O, Akman SA and Aksit S (2012) Vitamin D status of exclusively breastfed 4-month-old infants supplemented during different seasons. *Pediatrics* **130**:e921-927.

Hall P, Michels V, Gavrilov D, Matern D, Oglesbee D, Raymond K, Rinaldo P and Tortorelli S (2013) Aripiprazole and trazodone cause elevations of 7-dehydrocholesterol in the absence of Smith–Lemli–Opitz Syndrome. *Molecular Genetics and Metabolism* **110**:176-178.

Halldner L, Tillander A, Lundholm C, Boman M, Langstrom N, Larsson H and Lichtenstein P (2014) Relative immaturity and ADHD: findings from nationwide registers, parent- and self-reports. *Journal of child psychology and psychiatry, and allied disciplines* **55**:897-904.

Hallmann E, Lipowski J, Marszałek K and Rembiałkowska E (2013) The Seasonal Variation in Bioactive Compounds Content in Juice from Organic and Non-organic Tomatoes. *Plant Foods Hum Nutr* **68**:171-176.

Hao L, Ma J, Stampfer MJ, Ren A, Tian Y, Tang Y, Willett WC and Li Z (2003) Geographical, Seasonal and Gender Differences in Folate Status among Chinese Adults. *The Journal of Nutrition* **133**:3630-3635.

Hardy JB, McCracken GH, Jr, Gilkeson M and Sever JL (1969) Adverse fetal outcome following maternal rubella after the first trimes ter of pregnancy. *JAMA* **207**:2414-2420.

Hargraves J and Brennan N (2016) Medicare Hospice Spending Hit $15.8 Billion In 2015, Varied By Locale, Diagnosis. *Health Affairs* **35**:1902-1907.

Hay AW (1977) Tetrachlorodibenzo-p-dioxin release at Seveso. *Disasters* **1**:289-308.

Hayles AB and Nolan RB (1958) Masculinization of female fetus, possibly related to administration of progesterone during pregnancy; report of two cases. *Proceedings of the staff meetings Mayo Clinic* **33**:200-203.

Helsen WF, Van Winckel J and Williams AM (2005) The relative age effect in youth soccer across Europe. *Journal of sports sciences* **23**:629-636.

Henriksen JM (1986) Exercise-induced bronchoconstriction. Seasonal variation in children with asthma and in those with rhinitis. *Allergy* **41**:499-506.
Hernandez RK, Werler MM, Romitti P, Sun L and Anderka M (2012) Nonsteroidal antiinflammatory drug use among women and the risk of birth defects. *American Journal of Obstetrics and Gynecology* **206**:228.e221-228.e228.
HGMD (2015) Human Gene Mutation Database. http://www.hgmd.cf.ac.uk/ac/index.php. **Accessed in May 2015**.
Hill AB (1965) The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**:295.
Hill RM (1973) Drugs ingested by pregnant women. *Clinical Pharmacology & Therapeutics* **14**:654-659.
Hippocrates and Adams Ft (460BCE) On Airs, Waters, and Places. **<**http://classics.mit.edu/Hippocrates/airwatpl.mb.txt**>**
Hippocrates and Galen (1952) *Hippocratic Writings and On The Natural Faculties*, Encyclopaedia Britannica.
Hoffmann AA and Weeks AR (2007) Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in Drosophila melanogaster from eastern Australia. *Genetica* **129**:133-147.
Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H and Rabadan R (2011) Discovering disease associations by integrating electronic clinical data and medical literature. *PloS one* **6**:e21132.
Honein M, Paulozzi L and Erickson J (2001) Continued occurrence of Accutane®-exposed pregnancies. *Teratology* **64**:142-147.
Hopenhayn-Rich C, Biggs ML and Smith AH (1998) Lung and kidney cancer mortality associated with arsenic in drinking water in Cordoba, Argentina. *International Journal of Epidemiology* **27**:561-569.
Hopkins AL and Groom CR (2002) The druggable genome. *Nature reviews Drug discovery* **1**:727-730.
Hove-Madsen L, Llach A, Bayes-Genís A, Roura S, Font ER, Arís A and Cinca J (2004) Atrial Fibrillation Is Associated With Increased Spontaneous Calcium Release From the Sarcoplasmic Reticulum in Human Atrial Myocytes. *Circulation* **110**:1358-1363.
Hripcsak G and Albers DJ (2013) Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association : JAMIA* **20**:e311-318.
Hripcsak G, Knirsch C, Zhou L, Wilcox A and Melton G (2011) Bias associated with mining electronic health records. *Journal of biomedical discovery and collaboration* **6**:48-52.
Hripcsak G, Knirsch C, Zhou L, Wilcox A and Melton GB (2007) Using discordance to improve classification in narrative clinical databases: An application to community-acquired pneumonia. *Computers in Biology and Medicine* **37**:296-304.
Hsieh W-S, Wu H-C, Jeng S-F, Liao H-F, Su Y-N, Lin S-J, Hsieh C-J and Chen P-C (2006) Nationwide singleton birth weight percentiles by gestational age in Taiwan, 1998-2002. *Acta Paediatrica Taiwanica* **47**:25.
Huang DW, Sherman BT and Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols* **4**:44-57.

Huang DW, Sherman BT and Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**:1-13.
Huber S, Didham R and Fieder M (2008) Month of birth and offspring count of women: data from the Southern hemisphere. *Hum Reprod* **23**:1187-1192.
Huber S and Fieder M (2009) Strong association between birth month and reproductive performance of Vietnamese women. *American journal of human biology : the official journal of the Human Biology Council* **21**:25-35.
Huber S and Fieder M (2011) Perinatal winter conditions affect later reproductive performance in Romanian women: intra and intergenerational effects. *American journal of human biology : the official journal of the Human Biology Council* **23**:546-552.
Huber S, Fieder M, Wallner B, Moser G and Arnold W (2004) Brief communication: birth month influences reproductive performance in contemporary women. *Hum Reprod* **19**:1081-1082.
Hundhausen C, Boesch-Saadatmandi C, Matzner N, Lang F, Blank R, Wolffram S, Blaschek W and Rimbach G (2008) Ochratoxin A Lowers mRNA Levels of Genes Encoding for Key Proteins of Liver Cell Metabolism. *Cancer Genomics - Proteomics* **5**:319-332.
Ioos S, Mallet H-P, Goffart IL, Gauthier V, Cardoso T and Herida M (2014) Current Zika virus epidemiology and recent epidemics. *Medecine et maladies infectieuses* **44**:302-307.
IRS (2015) SOI Tax Stats - Individual Income Tax Statistics - 2012 ZIP Code Data (SOI). *Accessed on July 2015* <https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-2012-zip-code-data-soi>.
Isaacs R, Hunter W and Clark K (2001) Tamoxifen as systemic treatment of advanced breast cancer during pregnancy—case report and literature review. *Gynecologic oncology* **80**:405-408.
Ito K, Mathes R, Ross Z, Nádas A, Thurston G and Matte T (2011) Fine particulate matter constituents associated with cardiovascular hospitalizations and mortality in New York City. *Environmental health perspectives* **119**:467.
Jacob P, Goulko G, Heidenreich W, Likhtarev I, Kairo I, Tronko N, Bogdanova T, Kenigsberg J, Buglova E and Drozdovitch V (1998) Thyroid cancer risk to children calculated. *Nature* **392**:31-32.
Jain J, Samal B, Singhakowinta A and Vaitkevicius VK (1977) Clinical trial of nafoxidine in adrenalectomized patients with advanced breast cancer. *Cancer* **40**:2063-2066.
Jarup L (2004) Health and environment information systems for exposure and disease mapping, and risk assessment. *Environmental health perspectives*:995-997.
Jensen PB, Jensen LJ and Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* **13**:395-405.
Jezela-Stanek A, Ciara E, Małunowicz E, Chrzanowska K, Latos-Bieleńska A, Krajewska-Walasek M and Group S-L-OsC (2010) Differences between predicted and established diagnoses of Smith-Lemli-Opitz syndrome in the Polish population: underdiagnosis or loss of affected fetuses? *J Inherit Metab Dis* **33**:241-248.
Jha AK (2010) Meaningful use of electronic health records: the road ahead. *JAMA* **304**:1709-1710.
Jira P, Wanders R, Smeitink J, De Jong J, Wevers R, Oostheim W, Tuerlings J, Hennekam R, Sengers R and Waterham H (2001) Novel mutations in the 7-dehydrocholesterol

reductase gene of 13 patients with Smith–Lemli–Opitz syndrome. *Annals of human genetics* **65**:229-236.
Jira PE, Waterham HR, Wanders RJA, Smeitink JAM, Sengers RCA and Wevers RA (2003) Smith-Lemli-Opitz Syndrome and the DHCR7 Gene. *Annals of Human Genetics* **67**:269-280.
Joiner Jr TE, Pfaff JJ, Acres JG and Johnson F (2002) Birth month and suicidal and depressive symptoms in Australians born in the Southern vs. the Northern hemisphere. *Psychiatry Research* **112**:89-92.
Jordan VC (2003) Antiestrogens and selective estrogen receptor modulators as multifunctional medicines. 1. Receptor interactions. *Journal of medicinal chemistry* **46**:883-908.
Kahn CN, Ault T, Isenstein H, Potetz L and Van Gelder S (2006) Snapshot of hospital quality reporting and pay-for-performance under Medicare. *Health Affairs* **25**:148-162.
Kahn HS, Morgan TM, Case LD, Dabelea D, Mayer-Davis EJ, Lawrence JM, Marcovina SM and Imperatore G (2009) Association of Type 1 Diabetes With Month of Birth Among US Youth The SEARCH for Diabetes in Youth Study. *Diabetes Care* **32**:2010-2015.
Kangovi S and Grande D (2011) Hospital readmissions—not just a measure of quality. *JAMA* **306**:1796-1797.
Karakuła H, Szajer K, Rpila B, Grzywa A and Chuchra M (2004) Clozapine and pregnancy--a case history. *Pharmacopsychiatry* **37**:303-304.
Karliner LS, Kim SE, Meltzer DO and Auerbach AD (2010) Influence of language barriers on outcomes of hospital care for general medicine inpatients. *Journal of Hospital Medicine* **5**:276-282.
Karp G, Von Oeyen P, Valone F, Khetarpal V, Israel M, Mayer R, Frigoletto F and Garnick M (1983) Doxorubicin in pregnancy: possible transplacental passage. *Cancer treatment reports* **67**:773-777.
Karras S, Anagnostis P, Naughton D, Annweiler C, Petroczi A and Goulis D (2015) Vitamin D during pregnancy: why observational studies suggest deficiency and interventional studies show no improvement in clinical outcomes? A narrative review. *Journal of endocrinological investigation* **38**:1265-1275.
Kawata K, Yokoo H, Shimazaki R and Okabe S (2007) Classification of Heavy-Metal Toxicity by Human DNA Microarray Analysis. *Environmental Science & Technology* **41**:3769-3774.
Kay KM and Sargent RD (2009) The role of animal pollination in plant speciation: integrating ecology, geography, and genetics. *Annual Review of Ecology, Evolution, and Systematics* **40**:637-656.
Keller TH, Pichota A and Yin Z (2006) A practical view of 'druggability'. *Current Opinion in Chemical Biology* **10**:357-361.
Kelley RI and Hennekam RCM (2000) The Smith-Lemli-Opitz syndrome. *J Med Genet* **37**:321-335.
Kemkes A (2010) The impact of maternal birth month on reproductive performance: controlling for socio-demographic confounders. *Journal of biosocial science* **42**:177-194.
Kerr JR (2005) Neonatal Effects of Breast Cancer Chemotherapy Administered During Pregnancy. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **25**:438-441.

Khan MA and Khan MD (2005) Classification of 154 clinical cases of vitamin A deficiency in children (0-15 years) in a tertiary hospital in North West Frontier Province Pakistan. *J Pak Med Assoc* **55**:77-78.
Kidd BA, Wroblewska A, Boland MR, Agudo J, Merad M, Tatonetti NP, Brown BD and Dudley JT (2016) Mapping the effects of drugs on the immune system. *Nat Biotech* **34**:47-54.
Kim JH and Scialli AR (2011) Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicological Sciences* **122**:1-6.
Klink M, Bednarska K, Blus E, Kielbik M and Sulowska Z (2012) Seasonal changes in activities of human neutrophils in vitro. *Inflammation Research* **61**:11-16.
Knobeloch L and Jackson R (1999) Recognition of chronic carbon monoxide poisoning. *Wis Med J* **98**:26-29.
Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* **12**:417-428.
Koizumi K and Aono T (1986) Pregnancy after combined treatment with bromocriptine and tamoxifen in two patients with pituitary prolactinomas. *Fertility and sterility* **46**:312-314.
Köppen W (1884) The thermal zones of the Earth according to the duration of hot, moderate and cold periods and of the impact of heat on the organic world. *Meteorol Z* **20**:351-360.
Korsgaard J and Dahl R (1983) Sensitivity to house dust mite and grass pollen in adults. *Clinical & Experimental Allergy* **13**:529-536.
Kottek M (2015) World Map of the Koppen-Geiger Climate Classification Updated Map for the United States of America. http://koeppen-geiger.vu-wien.ac.at/data/KoeppenGeiger.UScounty.txt
Kottek M, Grieser J, Beck C, Rudolf B and Rubel F (2006) World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift* **15**:259-263.
Kozák L, Francová H, Hrabincová E, Procházková D, Jüttnerová V, Bzdúch V and Šimek P (2000) Smith–Lemli–Opitz syndrome: Molecular-genetic analysis of ten families. *J Inherit Metab Dis* **23**:409-412.
Kuan V, Martineau A, Griffiths C, Hypponen E and Walton R (2013) DHCR7 mutations linked to higher vitamin D status allowed early human migration to Northern latitudes. *BMC Evolutionary Biology* **13**:144.
Kulldorff M and Hjalmars U (1999) The Knox Method and Other Tests for Space-Time Interaction. *Biometrics* **55**:544-552.
Kumar J, Muntner P, Kaskel FJ, Hailpern SM and Melamed ML (2009) Prevalence and associations of 25-hydroxyvitamin D deficiency in US children: NHANES 2001-2004. *Pediatrics* **124**:e362-370.
Lacour-Gayet F, Clarke D, Jacobs J, Comas J, Daebritz S, Daenen W, Gaynor W, Hamilton L, Jacobs M, Maruszsewski B, Pozzi M, Spray T, Stellin G, Tchervenkov C and Mavroudis C (2004) The Aristotle score: a complexity-adjusted method to evaluate surgical results. *European Journal of Cardio-Thoracic Surgery* **25**:911-924.
Laland KN, Sterelny K, Odling-Smee J, Hoppitt W and Uller T (2011) Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? *science* **334**:1512-1516.
Laliberté E and Laliberté ME (2015) Package 'metacor'.
Lanphear BP, Hornung R, Khoury J, Yolton K, Baghurst P, Bellinger DC, Canfield RL, Dietrich KN, Bornschein R and Greene T (2005) Low-level environmental lead exposure and

children's intellectual function: an international pooled analysis. *Environmental health perspectives*:894-899.
Lanthaler B, Steichen-Gersdorf E, Kollerits B, Zschocke J and Witsch-Baumgartner M (2013) Maternal ABCA1 genotype is associated with severity of Smith–Lemli–Opitz syndrome and with viability of patients homozygous for null mutations. *European Journal of Human Genetics* **21**:286-293.
Lascano D, Finkelstein JB, Barlow LJ, Kabat D, RoyChoudhury A, Caso JR, DeCastro GJ, Gold W and McKiernan JM (2015) The Correlation of Media Ranking's "Best" Hospitals and Surgical Outcomes Following Radical Cystectomy for Urothelial Cancer. *Urology* **6**:1104-1114.
Lauth M, Rohnalter V, Bergstrom A, Kooshesh M, Svenningsson P and Toftgard R (2010) Antipsychotic drugs regulate hedgehog signaling by modulation of 7-dehydrocholesterol reductase levels. *Molecular pharmacology* **78**:486-496.
Lee CJ, Lawler GS and Panemangalore M (1987) Nutritional status of middle-aged and elderly females in Kentucky in two seasons: Part 2. Hematological parameters. *Journal of the American College of Nutrition* **6**:217-222.
Lee DM, Tajar A, Pye SR, Boonen S, Vanderschueren D, Bouillon R, O'Neill TW, Bartfai G, Casanueva FF and Finn JD (2012) Association of hypogonadism with vitamin D status: the European Male Ageing Study. *European Journal of Endocrinology* **166**:77-85.
Lee JH, O'Keefe JH, Bell D, Hensrud DD and Holick MF (2008) Vitamin D deficiency: an important, common, and easily treatable cardiovascular risk factor? *Journal of the American College of Cardiology* **52**:1949-1956.
Lee JM, Smith JR, Philipp BL, Chen TC, Mathieu J and Holick MF (2007) Vitamin D deficiency in a healthy group of mothers and newborn infants. *Clinical Pediatrics* **46**:42-44.
Lee KC, Korgavkar K, Dufresne RG and Higgins HW (2013) Safety of cosmetic dermatologic procedures during pregnancy. *Dermatologic Surgery* **39**:1573-1586.
Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A and Cummings B (2015) Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*:030338.
Lerchbaum E and Obermayer-Pietsch B (2012) Mechanisms in Endocrinology: Vitamin D and fertility: a systematic review. *European Journal of Endocrinology* **166**:765-778.
Levy O (2007) Innate immunity of the newborn: basic mechanisms and clinical correlates. *Nat Rev Immunol* **7**:379-390.
Li L, Boland M, Miotto R, Tatonetti NP and Dudley JT (2016) Replicating Cardiovascular Condition-Birth Month Associations. *Scientific Reports* **6**:33166.
Lim JW, Lee JJ, Park CG, Sriram S and Lee K-s (2010) Birth outcomes of Koreans by birthplace of infants and their mothers, the United States versus Korea, 1995-2004. *Journal of Korean medical science* **25**:1343-1351.
Liu CM (1988) Seasonal variation of nasal surface basophilic cells and eosinophils in Japanese cedar pollinosis. *Rhinology* **26**:167-173.
Liu H, Jacob DJ, Bey I and Yantosca RM (2001) Constraints from 210Pb and 7Be on wet deposition and transport in a global three‐dimensional chemical tracer model driven by assimilated meteorological fields. *Journal of Geophysical Research: Atmospheres* **106**:12109-12128.

Lorberbaum T, Nasir M, Keiser MJ, Vilar S, Hripcsak G and Tatonetti NP (2015) Systems Pharmacology Augments Drug Safety Surveillance. *Clinical Pharmacology & Therapeutics* **97**:151-158.

Loukides G, Gkoulalas-Divanis A and Malin B (2010) Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences* **107**:7898-7903.

Mactutus C and Fechter L (1984) Prenatal exposure to carbon monoxide: learning and memory deficits. *Science* **223**:409-411.

Mandel Y, Grotto I, El-Yaniv R, Belkin M, Israeli E, Polat U and Bartov E (2008) Season of Birth, Natural Light, and Myopia. *Ophthalmology* **115**:686-692.

Manson JM, Freyssinges C, Ducrocq MB and Stephenson WP (1996) Postmarketing surveillance of lovastatin and simvastatin exposure during pregnancy. *Reproductive Toxicology* **10**:439-446.

Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Guyer M and Green ED (2014) The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* **21**:957-958.

Marinoni A, Dagliati A, Bellazzi R and Gamba P (2015) Inferring air quality maps from remotely sensed data to exploit georeferenced clinical onsets: The Pavia 2013 case, in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International* pp 3937-3940, IEEE.

Maroziene L and Grazuleviciene R (2002) Maternal exposure to low-level air pollution and pregnancy outcomes: a population-based study. *Environmental Health* **1**:1.

Matsumoto Y, Morishima K-i, Honda A, Watabe S, Yamamoto M, Hara M, Hasui M, Saito C, Takayanagi T and Yamanaka T (2005) R352Q mutation of the DHCR7 gene is common among Japanese Smith–Lemli–Opitz syndrome patients. *Journal of human genetics* **50**:353-356.

Mazumder B, Almond D, Park K, Crimmins EM and Finch CE (2010) Lingering prenatal effects of the 1918 influenza pandemic on cardiovascular disease. *Journal of developmental origins of health and disease* **1**:26-34.

McGrath JJ, Eyles DW, Pedersen CB and et al. (2010) Neonatal vitamin d status and risk of schizophrenia: A population-based case-control study. *Archives of General Psychiatry* **67**:889-894.

McKenna K, Koren G, Tetelbaum M, Wilton L, Shakir S, Diav-Citrin O, Levinson A, Zipursky RB and Einarson A (2005) Pregnancy outcome of women using atypical antipsychotic drugs: a prospective comparative study. *The Journal of clinical psychiatry* **66**:444-449; quiz 546.

McKinley MC, Strain JJ, McPartlin J, Scott JM and McNulty H (2001) Plasma Homocysteine Is Not Subject to Seasonal Variation. *Clinical Chemistry* **47**:1430-1436.

Megías-Vericat J, Rojas L, Herrero M, Bosó V, Montesinos P, Moscardó F, Poveda J, Sanz MÁ and Aliño S (2015) Influence of ABCB1 polymorphisms upon the effectiveness of standard treatment for acute myeloid leukemia: A systematic review and meta-analysis of observational studies. *The pharmacogenomics journal*.

Meier C, Woitge HW, Witte K, Lemmer B and Seibel MJ (2004) Supplementation with oral vitamin D3 and calcium during winter prevents seasonal bone loss: a randomized controlled open-label prospective trial. *Journal of Bone and Mineral Research* **19**:1221-1230.

Melamed RD, Emmett KJ, Madubata C, Rzhetsky A and Rabadan R (2015) Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes. *Nat Commun* **6**.
Melamed RD, Khiabanian H and Rabadan R (2014) Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. *BMC bioinformatics* **15 Suppl 6**:S3.
Melnikov V, Suvorova IY and Belisheva N (2016) Central hemodynamics and arterial stiffness in adult humans depend on the conditions of early development in the Northern Kola Peninsula. *Human Physiology* **42**:150-155.
Melnikov VN (2003) Life span of people who died from cardiovascular diseases in Siberia: a comparative study of two populations. *International journal of circumpolar health* **62**.
Melnikov VN, Skosyreva GA and Krivoschekov SG (2007) Seasonality bias in adverse pregnancy outcomes in Siberia. *Alaska Med* **49**:218-220.
Mendhekar D and Andrade C (2011) Uneventful use of haloperidol and trihehexyphenidyl during three consecutive pregnancies. *Archives of women's mental health* **14**:83-84.
Mercier G, Georgescu V and Bousquet J (2015) Geographic variation in potentially avoidable hospitalizations in France. *Health Affairs* **34**:836-843.
Mereu G, Cammalleri M, Fà M, Francesconi W, Saba P, Tattoli M, Trabace L, Vaccari A and Cuomo V (2000) Prenatal Exposure to a Low Concentration of Carbon Monoxide Disrupts Hippocampal Long-Term Potentiation in Rat Offspring. *Journal of Pharmacology and Experimental Therapeutics* **294**:728-734.
Meyer-Wittkopf M, Barth H, Emons G and Schmidt S (2001) Fetal cardiac effects of doxorubicin therapy for carcinoma of the breast during pregnancy: case report and review of the literature. *Ultrasound in Obstetrics and Gynecology* **18**:62-66.
Miettola S, Hartikainen A-L, Vääräsmäki M, Bloigu A, Ruokonen A, Järvelin M-R and Pouta A (2013) Offspring's blood pressure and metabolic phenotype after exposure to gestational hypertension in utero. *European journal of epidemiology* **28**:87-98.
Miliku K, Voortman T, Franco OH, McGrath JJ, Eyles DW, Burne TH, Hofman A, Tiemeier H and Jaddoe VWV (2015) Vitamin D status during fetal life and childhood kidney outcomes. *Eur J Clin Nutr*.
Milton AH, Smith W, Rahman B, Hasan Z, Kulsum U, Dear K, Rakibuddin M and Ali A (2005) Chronic Arsenic Exposure and Adverse Pregnancy Outcomes in Bangladesh. *Epidemiology* **16**:82-86.
Mir O, Berrada N, Domont J, Cioffi A, Boulet B, Terrier P, Bonvalot S, Trichot C, Lokiec F and Le Cesne A (2012) Doxorubicin and ifosfamide for high-grade sarcoma during pregnancy. *Cancer chemotherapy and pharmacology* **69**:357-367.
Mocarelli P, Brambilla P, Gerthoux PM, Patterson Jr DG and Needham LL (1996) Change in sex ratio with exposure to dioxin. *The Lancet* **348**:409.
Mocarelli P, Gerthoux PM, Ferrari E, Patterson DG, Kieszak SM, Brambilla P, Vincoli N, Signorini S, Tramacere P and Carreri V (2000) Paternal concentrations of dioxin and sex ratio of offspring. *The Lancet* **355**:1858-1863.
Moebius FF, Fitzky BU, Lee JN, Paik YK and Glossmann H (1998) Molecular cloning and expression of the human delta7-sterol reductase. *Proceedings of the National Academy of Sciences of the United States of America* **95**:1899-1902.
Moffitt TE, Harrington H, Caspi A and et al. (2007) Depression and generalized anxiety disorder: Cumulative and sequential comorbidity in a birth cohort followed prospectively to age 32 years. *Archives of General Psychiatry* **64**:651-660.

Møller AP and Mousseau TA (2006) Biological consequences of Chernobyl: 20 years on. *Trends in ecology & evolution* **21**:200-207.
Møller AP, Surai P and Mousseau T (2005) Antioxidants, radiation and mutation as revealed by sperm abnormality in barn swallows from Chernobyl. *Proceedings of the Royal Society of London B: Biological Sciences* **272**:247-253.
Mora JR, Iwata M and von Andrian UH (2008) Vitamin effects on the immune system: vitamins A and D take centre stage. *Nat Rev Immunol* **8**:685-698.
Morita Y and Tilly JL (1999) Oocyte apoptosis: like sand through an hourglass. *Developmental biology* **213**:1-17.
Morris PG, King F and Kennedy MJ (2009) Cytotoxic chemotherapy for pregnancy-associated breast cancer: single institution case series. *Journal of Oncology Pharmacy Practice* **15**:241-247.
Mühlenweg AM (2010) Young and innocent: international evidence on age effects within grades on victimization in elementary school. *Economics Letters* **109**:157-160.
Murray CL, Reichert JA, Anderson J and Twiggs LB (1984) Multimodal cancer therapy for breast cancer in the first trimester of pregnancy: A case report. *Jama* **252**:2607-2608.
Musch J and Grondin S (2001) Unequal competition as an impediment to personal development: A review of the relative age effect in sport. *Developmental review* **21**:147-167.
Naeye RL and Blanc W (1965) Pathogenesis of congenital rubella. *JAMA* **194**:1277-1283.
Navas-Acien A, Guallar E, Silbergeld EK and Rothenberg SJ (2007) Lead exposure and cardiovascular disease: a systematic review. *Environmental health perspectives*:472-482.
Navas-Acien A, Sharrett AR, Silbergeld EK, Schwartz BS, Nachman KE, Burke TA and Guallar E (2005) Arsenic Exposure and Cardiovascular Disease: A Systematic Review of the Epidemiologic Evidence. *American Journal of Epidemiology* **162**:1037-1049.
NCES (2016) Children and Youth with Disabilities *Accessed on May 24, 2016* <http://nces.ed.gov/programs/coe/indicator_cgg.asp>.
Nelissen ECM, van Montfoort APA, Dumoulin JCM and Evers JLH (2011) Epigenetics and the placenta. *Human Reproduction Update* **17**:397-417.
Neto LV, De Almeida CA, Da Costa SM and Vaisman M (2007) Prospective evaluation of pregnant women with hypothyroidism: Implications for treatment. *Gynecological endocrinology* **23**:138-141.
Newport DJ, Calamaras MR, DeVane CL, Donovan J, Beach AJ, Winn S, Knight BT, Gibson BB, Viguera AC and Owens MJ (2007) Atypical antipsychotic administration during late pregnancy: placental passage and obstetrical outcomes.
Nezarati MM, Loeffler J, Yoon G, MacLaren L, Fung E, Snyder F, Utermann G and Graham GE (2002) Novel mutation in the Δ-sterol reductase gene in three Lebanese sibs with Smith-Lemli-Opitz (RSH) syndrome. *American journal of medical genetics* **110**:103-108.
Nielsen GL, Sorensen HT, Larsen H and Pedersen L (2001) Risk of adverse birth outcome and miscarriage in pregnant users of non-steroidal anti-inflammatory drugs: population based observational study and case-control study. *Bmj* **322**:266-270.
Nieto Y, Santisteban M, Aramendía JM, Fernández-Hidalgo Ó, García-Manero M and López G (2006) Docetaxel administered during pregnancy for inflammatory breast carcinoma. *Clinical breast cancer* **6**:533-534.
Norman AW (1998) Sunlight, season, skin pigmentation, vitamin D, and 25-hydroxyvitamin D: integral components of the vitamin D endocrine system. *American Journal of Clinical Nutrition* **67**:1108-1110.

Nowaczyk MJM and Irons MB (2012) Smith–Lemli–Opitz syndrome: Phenotype, natural history, and epidemiology. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **160C**:250-262.

Nowaczyk MJM, Martin-Garcia D, Aquino-Perna A, Rodriguez-Vazquez M, McCaughey D, Eng B, Nakamura LM and Waye JS (2004a) Founder effect for the T93M DHCR7 mutation in Smith-Lemli-Opitz syndrome. *American Journal of Medical Genetics Part A* **125A**:173-176.

Nowaczyk MJM, Nakamura LM, Eng B, Porter FD and Waye JS (2001) Frequency and ethnic distribution of the common DHCR7 mutation in Smith-Lemli-Opitz syndrome. *American Journal of Medical Genetics* **102**:383-386.

Nowaczyk MJM, Waye JS and Douketis JD (2006) DHCR7 mutation carrier rates and prevalence of the RSH/Smith-Lemli-Opitz syndrome: Where are the patients? *American Journal of Medical Genetics Part A* **140A**:2057-2062.

Nowaczyk MJM, Zeesman S, Waye JS and Douketis JD (2004b) Incidence of Smith-Lemli-Opitz syndrome in Canada: Results of three-year population surveillance. *The Journal of Pediatrics* **145**:530-535.

NYSDOH (2007) Congenital Malformations Registry - Summary Report. Appendix 1: Classification of Codes. <https://wwwhealthnygov/diseases/congenital_malformations/2002_2004/appendiceshtm > **Accessed on 11/30/2016**.

Ofori B, Oraichi D, Blais L, Rey E and Bérard A (2006) Risk of congenital anomalies in pregnant users of non-steroidal anti-inflammatory drugs: A nested case-control study. *Birth Defects Research Part B: Developmental and Reproductive Toxicology* **77**:268-279.

Oh M-Y, Kim JS, Kim JH, Cho JH, Lee BH, Kim G-H, Choi J-H and Yoo H-W (2014) A case of Smith-Lemli-Opitz syndrome confirmed by molecular analysis: Review of mutation spectrum of the DHCR7 gene in Korea. *Journal of Genetic Medicine* **11**:106-110.

Öksüzoglu B and Güler N (2002) An infertile patient with breast cancer who delivered a healthy child under adjuvant tamoxifen therapy. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **104**:79.

Olesen C, Hald Steffensen F, Lauge Nielsen G, Jong-van den Berg L, Olsen J and Toft Sørensen H (1999) Drug use in first pregnancy and lactation: a population-based survey among Danish women. *European journal of clinical pharmacology* **55**:139-144.

Opitz JM, Gilbert-Barness E, Ackerman J and Lowichik A (2002) Cholesterol and development: the RSH (" Smith-Lemli-Opitz") syndrome and related conditions. *Fetal & Pediatric Pathology* **21**:153-181.

Otake M, Yoshimaru H and Schull WJ (1988) Severe mental retardation among the prenatally exposed survivors of the Atomic bombing of Hiroshima and Nagasaki, Radiation Effects Research Foundation.

Overhage JM, Ryan PB, Reich CG, Hartzema AG and Stang PE (2012) Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* **19**:54-60.

Ozkan S, Jindal S, Greenseid K, Shu J, Zeitlian G, Hickmon C and Pal L (2010) Replete vitamin D stores predict reproductive success following in vitro fertilization. *Fertility and Sterility* **94**:1314-1319.

Paalanen L, Prattala R, Alfthan G, Salminen I and Laatikainen T (2013) Seasonal variation in plasma vitamin C concentration in Pitkaranta, Northwestern Russia. *Eur J Clin Nutr* **67**:1115-1115.
Palva I and Salokannel S (1972) Seasonal variation in megaloblastic anaemia. *British Journal of Nutrition* **27**:593-595.
Pantazatos SP (2014) Prediction of individual season of birth using MRI. *NeuroImage* **88**:61-68.
Park E-J, Zahari NEM, Lee E-W, Song J, Lee J-H, Cho M-H and Kim J-H (2014) SWCNTs induced autophagic cell death in human bronchial epithelial cells. *Toxicology in Vitro* **28**:442-450.
Parks JH, Barsky R and Coe FL (2003) Gender differences in seasonal variation of urine stone risk factors. *The Journal of urology* **170**:384-388.
Patel CJ, Bhattacharya J and Butte AJ (2010) An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS ONE* **5**:e10746.
Patrono C, Dionisi-Vici C, Giannotti A, Bembi B, Digilio M, Rizzo C, Purificato C, Martini C, Pierini R and Santorelli F (2002) Two novel mutations of the human Δ7-sterol reductase (DHCR7) gene in children with Smith–Lemli–Opitz syndrome. *Molecular and cellular probes* **16**:315-318.
Peccatori F, Martinelli G, Gentilini O and Goldhirsch A (2004) Chemotherapy during pregnancy: what is really safe? *The Lancet Oncology* **5**:398.
Peel MC, Finlayson BL and McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* **11**:1633-1644.
Percy ME, Andrews DF and Thompson MW (1982) Serum creatine kinase in the detection of Duchenne muscular dystrophy carriers: effects of season and multiple testing. *Muscle & nerve* **5**:58-64.
Peterka M, Peterková R and Likovský Z (2004) Chernobyl: prenatal loss of four hundred male fetuses in the Czech Republic. *Reproductive Toxicology* **18**:75-79.
Peterson HdC (1960) Acquired methemoglobinemia in an infant due to benzocaine suppository. *New England Journal of Medicine* **263**:454-455.
Philbin EF, Dec GW, Jenkins PL and DiSalvo TG (2001) Socioeconomic status as an independent risk factor for hospital readmission for heart failure. *The American Journal of Cardiology* **87**:1367-1371.
Piazza A, Menozzi P and Cavalli-Sforza LL (1981) Synthetic gene frequency maps of man and selective effects of climate. *Proceedings of the National Academy of Sciences* **78**:2638-2642.
Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F and Furlong LI (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**.
Polack S, Brooker S, Kuper H, Mariotti S, Mabey D and Foster A (2005) Mapping the global distribution of trachoma. *Bulletin of the World Health Organization* **83**:913-919.
Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM and Posthuma D (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**:702-709.
Pollack PS, Shields KE, Burnett DM, Osborne MJ, Cunningham ML and Stepanavage ME (2005) Pregnancy outcomes after maternal exposure to simvastatin and lovastatin. *Birth Defects Research Part A: Clinical and Molecular Teratology* **73**:888-896.

Pomorski L, Bartos M and Narebski J (1999) Pregnancy following operative and complementary treatment of thyroid cancer. *Zentralblatt fur Gynakologie* **122**:383-386.
Porlier M, Bélisle M and Garant D (2009) Non-random distribution of individual genetic diversity along an environmental gradient. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**:1543-1554.
Porter FD (2008) Smith-Lemli-Opitz syndrome: pathogenesis, diagnosis and management. *European journal of human genetics : EJHG* **16**:535-541.
Pöschl U (2005) Atmospheric aerosols: composition, transformation, climate and health effects. *Angewandte Chemie International Edition* **44**:7520-7540.
Potluri V, Lewis D and Burton GV (2006) Chemotherapy with taxanes in breast cancer during pregnancy: case report and review of the literature. *Clinical breast cancer* **7**:167-170.
Prasad C, Marles S, Prasad AN, Nikkel S, Longstaffe S, Peabody D, Eng B, Wright S, Waye JS and Nowaczyk MJM (2002) Smith-Lemli-Opitz syndrome: New mutation with a mild phenotype. *American Journal of Medical Genetics* **108**:64-68.
Psaty BM, Manolio TA, Kuller LH, Kronmal RA, Cushman M, Fried LP, White R, Furberg CD and Rautaharju PM (1997) Incidence of and risk factors for atrial fibrillation in older adults. *Circulation* **96**:2455-2461.
Raeder MB, Fernø J, Vik-Mo AO and Steen VM (2006) SREBP activation by antipsychotic-and antidepressant-drugs in cultured human liver cells: relevance for metabolic side-effects? *Molecular and cellular biochemistry* **289**:167-173.
Rahman A, Vahter M, Smith AH, Nermell B, Yunus M, El Arifeen S, Persson L-Å and Ekström E-C (2009) Arsenic Exposure During Pregnancy and Size at Birth: A Prospective Cohort Study in Bangladesh. *American Journal of Epidemiology* **169**:304-312.
Ramey JA (2015) U.S. Census Regional and Demographic Data. *Package 'noncensus'* https://cran.r-project.org/web/packages/noncensus/noncensus.pdf
Randolph C (2014) Seasonality of asthma: a retrospective population study. *Pediatrics* **134 Suppl 3**:S165-166.
Rasmussen SA, Jamieson DJ, Honein MA and Petersen LR (2016) Zika Virus and Birth Defects — Reviewing the Evidence for Causality. *New England Journal of Medicine* **374**:1981-1987.
Raub JA, Mathieu-Nolf M, Hampson NB and Thom SR (2000) Carbon monoxide poisoning—a public health perspective. *Toxicology* **145**:1-14.
Reis JP, von Muhlen D, Miller ER, 3rd, Michos ED and Appel LJ (2009) Vitamin D status and cardiometabolic risk factors in the United States adolescent population. *Pediatrics* **124**:e371-379.
Reis M and Källén B (2008) Maternal Use of Antipsychotics in Early Pregnancy and Delivery Outcome. *Journal of Clinical Psychopharmacology* **28**:279-288.
Roach E, Demyer W, Conneally P, Palmer C and Merritt A (1974) Holoprosencephaly: birth data, genetic and demographic analyses of 30 families. *Birth defects original article series* **11**:294-313.
Rodgers MA, Villareal VA, Schaefer EA, Peng LF, Corey KE, Chung RT and Yang PL (2012) Lipid Metabolite Profiling Identifies Desmosterol Metabolism as a New Antiviral Target for Hepatitis C Virus. *Journal of the American Chemical Society* **134**:6896-6899.
Rogers JF, Thompson SJ, Addy CL, McKeown RE, Cowen DJ and Decoufle P (2000) Association of very low birth weight with exposures to environmental sulfur dioxide and total suspended particulates. *American Journal of Epidemiology* **151**:602-613.

Romano F, Fiore B, Pezzino FM, Longombardo MT, Cefalù AB, Noto D, Puglisi A, Brogna A, Mattina T and Averna M (2005) A novel mutation of the DHCR7 gene in a Sicilian compound heterozygote with Smith-Lemli-Opitz syndrome. *Molecular Diagnosis* **9**:201-204.

Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A and Werge T (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology* **7**:e1002141.

Rossi M, Federico G, Corso G, Parenti G, Battagliese A, Frascogna AR, della Casa R, Dello Russo A, Strisciuglio P, Saggese G and Andria G (2005) Vitamin D status in patients affected by Smith-Lemli-Opitz syndrome. *J Inherit Metab Dis* **28**:69-80.

Roth DE (2011) Vitamin D supplementation during pregnancy: safety considerations in the design and interpretation of clinical trials. *J Perinatol* **31**:449-459.

Rotondi M, Caccavale C, Di Serio C, Del Buono A, Sorvillo F, Glinoer D, Bellastella A and Carella C (1999) Successful outcome of pregnancy in a thyroidectomized-parathyroidectomized young woman affected by severe hypothyroidism. *Thyroid* **9**:1037-1040.

Rudolph AJ, Yow MD, Phillips CA, Desmond MM, Blattner RJ and Melnick JL (1965) Transplacental rubella infection in newly born infants. *JAMA* **191**:843-845.

Ruenitz PC, Arrendale RF, Schmidt WF, Thompson CB and Nanavati NT (1989) Phenolic metabolites of clomiphene: [(E,Z)-2-[4-(1,2-diphenyl-2-chlorovinyl)phenoxy]ethyl]diethylamine. Preparation, electrophilicity, and effects in MCF 7 breast cancer cells. *Journal of medicinal chemistry* **32**:192-197.

Salam MT, Millstein J, Li Y-F, Lurmann FW, Margolis HG and Gilliland FD (2005) Birth outcomes and prenatal exposure to ozone, carbon monoxide, and particulate matter: results from the Children's Health Study. *Environmental health perspectives*:1638-1644.

Sapolsky RM (2001) Depression, antidepressants, and the shrinking hippocampus. *Proceedings of the National Academy of Sciences* **98**:12320-12322.

Sawhney H, Vasishta K and Rani K (1998) Comparison of lytic cocktail and magnesium sulphate regimens in eclampsia: a retrospective analysis. *Journal of Obstetrics and Gynaecology Research* **24**:261-266.

Scalco F, Correa-Cerro L, Wassif C, Porter F and Moretti-Ferreira D (2005) DHCR7 mutations in Brazilian Smith–Lemli–Opitz syndrome patients. *American Journal of Medical Genetics Part A* **136**:278-281.

Schaefer C, Meister R and Weber-Schoendorfer C (2010) Isotretinoin exposure and pregnancy outcome: an observational study of the Berlin Institute for Clinical Teratology and Drug Risk Assessment in Pregnancy. *Archives of gynecology and obstetrics* **281**:221-227.

Schaff EA, Eisinger SH, Stadalius LS, Franks P, Gore BZ and Poppema S (1999) Low-dose mifepristone 200 mg and vaginal misoprostol for abortion. *Contraception* **59**:1-6.

Schemske DW, Mittelbach GG, Cornell HV, Sobel JM and Roy K (2009) Is there a latitudinal gradient in the importance of biotic interactions? *Annu Rev Ecol Evol Syst* **40**:245-269.

Scherb H and Voigt K (2007) Trends in the human sex odds at birth in Europe and the Chernobyl Nuclear Power Plant accident. *Reproductive Toxicology* **23**:593-599.

Schull WJ, Otake M and Yoshimaru H (1988) Effect on intelligence test score of prenatal exposure to ionizing radiation in Hiroshima and Nagasaki, Radiation Effects Research Foundation.

Schulze R (2004) *Meta-analysis-A comparison of approaches*, Hogrefe Publishing.

Schwartz J (1999) Air Pollution and Hospital Admissions for Heart Disease in Eight U.S. Counties. *Epidemiology* **10**:17-22.
Schwartz J, Samet JM and Patz JA (2004) Hospital Admissions for Heart Disease: The Effects of Temperature and Humidity. *Epidemiology* **15**:755-761.
Scott-Phillips TC, Dickins TE and West SA (2011) Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science* **6**:38-47.
Secnik K, Swensen A and Lage M (2005) Comorbidities and costs of adult patients diagnosed with attention-deficit hyperactivity disorder. *Pharmacoeconomics* **23**:93-102.
Sever JL, Nelson KB and Gilkeson M (1965) Rubella epidemic, 1964: Effect on 6,000 pregnancies: i. preliminary clinical and laboratory findings through the neonatal period: a report from the collaborative study on cerebral palsy. *American Journal of Diseases of Children* **110**:395-407.
Shah NH (2013) Mining the ultimate phenome repository. *Nat Biotech* **31**:1095-1097.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**:2498-2504.
Sharma AP, Saeed A, Durani S and Kapil RS (1990) Structure-activity relationship of antiestrogens. Phenolic analogues of 2,3-diaryl-2H-1-benzopyrans. *Journal of medicinal chemistry* **33**:3222-3229.
Sherwood SC and Huber M (2010) An adaptability limit to climate change due to heat stress. *Proceedings of the National Academy of Sciences* **107**:9552-9555.
Shim Y-H, Bae S-H, Kim J-H, Kim K-R, Kim CJ and Paik Y-K (2004) A novel mutation of the human 7-dehydrocholesterol reductase gene reduces enzyme activity in patients with holoprosencephaly. *Biochemical and Biophysical Research Communications* **315**:219-223.
Singh J (1987) Nitrogen dioxide exposure alters neonatal development. *Neurotoxicology* **9**:545-549.
Smith DW, Lemli L and Opitz JM (1964) A newly recognized syndrome of multiple congenital anomalies. *J Pediatr* **64**:210-217.
Smith GD and Ebrahim S (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology* **32**:1-22.
Smith TM, Shugart HH, Bonan GB and Smith JB (1992) Modeling the Potential Response of Vegetation to Global Climate Change, in *Advances in Ecological Research* (M. Begon AHF and Macfadyen A eds) pp 93-116, Academic Press.
Smithells R (1962) Thalidomide and malformations in Liverpool. *The Lancet* **279**:1270-1273.
Soliman KB, Abbas MM, Seksaka MA, Wafa S and Balah AS (2007) Aggressive primary thyroid non Hodgkin's lymphoma with pregnancy. *Saudi medical journal* **28**:634-636.
Stamler J, Stamler R, Neaton JD and et al. (1999) Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy: Findings for 5 large cohorts of young adult and middle-aged men and women. *JAMA* **282**:2012-2018.
Steinbaum FL, De Jager RL and Krakoff IH (1978) Clinical trial of nafoxidine in advanced breast cancer. *Medical and pediatric oncology* **4**:123-126.

Steingrimsdottir L, Gunnarsson O, Indridason OS, Franzson L and Sigurdsson G (2005) Relationship between serum parathyroid hormone levels, vitamin d sufficiency, and calcium intake. *JAMA* **294**:2336-2341.

Steinhauser G, Brandl A and Johnson TE (2014) Comparison of the Chernobyl and Fukushima nuclear accidents: A review of the environmental impacts. *Science of The Total Environment* **470–471**:800-817.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD and Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**:1-9.

Stern JM and Simes RJ (1997) *Publication bias: evidence of delayed publication in a cohort study of clinical research projects*.

Stoner SC, Sommi R, Marken PA, Anya I and Vaughn J (1997) Clozapine use in two full-term pregnancies. *The Journal of clinical psychiatry* **58**:364-365.

Sunitha T, Rebekah Prasoona K, Muni Kumari T, Srinadh B, Laxmi Naga Deepika M, Aruna R and Jyothy A (2017) Risk factors for congenital anomalies in high risk pregnant women: A large study from South India. *Egyptian Journal of Medical Human Genetics* **18**:79-85.

Svedenhag J and Sjödin B (1985) Physiological characteristics of elite male runners in and off-season. *Canadian journal of applied sport sciences Journal canadien des sciences appliquees au sport* **10**:127-133.

Swan C, Tostevin A, Moore B, Mayo H and Black GB (1943) Congenital Defects in Infants following Infectious Diseases during Pregnancy. With special reference to the Relationship between German Measles and Cataract, Deaf-Mutism, Heart Disease and Microcephaly, and to the Period of Pregnancy in which the Occurrence of Rubella is followed by Congenital Abnormalities. *Medical journal of Australia* **2**:201-210.

Szabó G, Oláh A, Kozak L, Balogh E, Nagy A, Blahakova I and Oláh É (2010) A patient with Smith–Lemli–Opitz syndrome: novel mutation of the DHCR7 gene and effects of therapy with simvastatin and cholesterol supplement. *Eur J Pediatr* **169**:121-123.

Taguchi N, Rubin ET, Hosokawa A, Choi J, Ying AY, Moretti ME, Koren G and Ito S (2008) Prenatal exposure to HMG-CoA reductase inhibitors: effects on fetal and neonatal outcomes. *Reproductive Toxicology* **26**:175-177.

Tatonetti NP, Fernald GH and Altman RB (2012) A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association* **19**:79-85.

Taylor PN, Minassian C, Rehman A, Iqbal A, Draman MS, Hamilton W, Dunlop D, Robinson A, Vaidya B and Lazarus JH (2014) TSH levels and risk of miscarriage in women on long-term levothyroxine: a community-based study. *The Journal of Clinical Endocrinology & Metabolism* **99**:3895-3902.

Teo CC, Kon OL, Sim KY and Ng SC (1992) Synthesis of 2-(p-chlorobenzyl)-3-aryl-6-methoxybenzofurans as selective ligands for antiestrogen-binding sites. Effects on cell proliferation and cholesterol synthesis. *Journal of medicinal chemistry* **35**:1330-1339.

Tewari K, Bonebrake RG, Asrat T and Shanberg AM (1997) Ambiguous genitalia in infant exposed to tamoxifen in utero. *The Lancet* **350**:183.

Tilly JL, Niikura Y and Rueda BR (2009) The Current Status of Evidence for and Against Postnatal Oogenesis in Mammals: A Case of Ovarian Optimism Versus Pessimism? *Biology of Reproduction* **80**:2-12.

Tint GS, Irons M, Elias ER, Batta AK, Frieden R, Chen TS and Salen G (1994) Defective Cholesterol Biosynthesis Associated with the Smith-Lemli-Opitz Syndrome. *New England Journal of Medicine* **330**:107-113.

Trussell GC and Etter RJ (2001) Integrating genetic and environmental forces that shape the evolution of geographic variation in a marine snail, in *Microevolution Rate, Pattern, Process* pp 321-337, Springer.

VA (2015a) Benefits for Veterans' Children with Birth Defects. <http://wwwpublichealthvagov/exposures/agentorange/benefits/children-birth-defectsasp> **Accessed in November 2015**.

VA (2015b) Birth Defects in Children of Women Vietnam Veterans <http://wwwpublichealthvagov/exposures/agentorange/birth-defects/children-women-vietnam-vetsasp> **Accessed in November 2015**.

Vawdrey DK and Hripcsak G (2013) Publication bias in clinical trials of electronic health records. *Journal of Biomedical Informatics* **46**:139-141.

Ventura SJ, Curtin SC, Abma JC and Henshaw SK (2012) Estimated pregnancy rates and rates of pregnancy outcomes for the United States, 1990-2008. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* **60**:1-21.

Vinikoor-Imler LC, Gray SC, Edwards SE and Miranda ML (2012) The effects of exposure to particulate matter and neighbourhood deprivation on gestational hypertension. *Paediatric and perinatal epidemiology* **26**:91-100.

Waldie KE, Poulton R, Kirk IJ and Silva PA (2000) The effects of pre-and post-natal sunlight exposure on human growth: evidence from the Southern Hemisphere. *Early human development* **60**:35-42.

Wang TJ, Pencina MJ, Booth SL, Jacques PF, Ingelsson E, Lanier K, Benjamin EJ, D'Agostino RB, Wolf M and Vasan RS (2008) Vitamin D deficiency and risk of cardiovascular disease. *Circulation* **117**:503-511.

Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, Berry D, Kiel DP, Streeten EA, Ohlsson C, Koller DL, Peltonen L, Cooper JD, O'Reilly PF, Houston DK, Glazer NL, Vandenput L, Peacock M, Shi J, Rivadeneira F, McCarthy MI, Anneli P, de Boer IH, Mangino M, Kato B, Smyth DJ, Booth SL, Jacques PF, Burke GL, Goodarzi M, Cheung CL, Wolf M, Rice K, Goltzman D, Hidiroglou N, Ladouceur M, Wareham NJ, Hocking LJ, Hart D, Arden NK, Cooper C, Malik S, Fraser WD, Hartikainen AL, Zhai G, Macdonald HM, Forouhi NG, Loos RJ, Reid DM, Hakim A, Dennison E, Liu Y, Power C, Stevens HE, Jaana L, Vasan RS, Soranzo N, Bojunga J, Psaty BM, Lorentzon M, Foroud T, Harris TB, Hofman A, Jansson JO, Cauley JA, Uitterlinden AG, Gibson Q, Jarvelin MR, Karasik D, Siscovick DS, Econs MJ, Kritchevsky SB, Florez JC, Todd JA, Dupuis J, Hypponen E and Spector TD (2010) Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* **376**:180-188.

Wang X, Hripcsak G, Markatou M and Friedman C (2009) Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* **16**:328-337.

Wannamethee SG, Welsh P, Papacosta O, Lennon L, Whincup PH and Sattar N (2014) Elevated parathyroid hormone, but not vitamin D deficiency, is associated with increased risk of heart failure in older men with and without cardiovascular disease. *Circulation Heart failure* **7**:732-739.

Wassif CA, Krakowiak PA, Wright BS, Gewandter JS, Sterner AL, Javitt N, Yergey AL and Porter FD (2005) Residual cholesterol synthesis and simvastatin induction of cholesterol synthesis in Smith–Lemli–Opitz syndrome fibroblasts. *Molecular Genetics and Metabolism* **85**:96-107.

Waterham HR and Hennekam RCM (2012) Mutational spectrum of Smith–Lemli–Opitz syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **160C**:263-284.

Waterham HR and Wanders RJA (2000) Biochemical and genetic aspects of 7-dehydrocholesterol reductase and Smith-Lemli-Opitz syndrome. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1529**:340-356.

Waterham HR, Wijburg FA, Hennekam RCM, Vreken P, Poll-The BT, Dorland L, Duran M, Jira PE, Smeitink JAM, Wevers RA and Wanders RJA (1998) Smith-Lemli-Opitz Syndrome Is Caused by Mutations in the 7-Dehydrocholesterol Reductase Gene. *The American Journal of Human Genetics* **63**:329-338.

Waye J, Nakamura L, Eng B, Hunnisett L, Chitayat D, Costa T and Nowaczyk M (2002) Smith-Lemli-Opitz syndrome: carrier frequency and spectrum of DHCR7 mutations in Canada. *Journal of medical genetics* **39**:e31-e31.

Waye JS, Eng B, Potter MA, Nowaczyk MJ, McFadden D and Langlois S (2007) De novo mutation of the DHCR7 gene in a fetus with severe Smith–Lemli–Opitz (or RSH) syndrome. *American Journal of Medical Genetics Part A* **143**:1799-1801.

Waye JS, Krakowiak PA, Wassif CA, Sterner AL, Eng B, Nakamura LM, Nowaczyk MJM and Porter FD (2005) Identification of nine novel DHCR7 missense mutations in patients with Smith-Lemli-Opitz syndrome (SLOS). *Human Mutation* **26**:59-59.

Webb AR, Kline L and Holick MF (1988) Influence of Season and Latitude on the Cutaneous Synthesis of Vitamin D3: Exposure to Winter Sunlight in Boston and Edmonton Will Not Promote Vitamin D3 Synthesis in Human Skin. *The Journal of Clinical Endocrinology & Metabolism* **67**:373-378.

Webster WS (1998) Teratogen update: congenital rubella. *Teratology* **58**:13-23.

Wehr E, Trummer O, Giuliani A, Gruber H-J, Pieber TR and Obermayer-Pietsch B (2011) Vitamin D-associated polymorphisms are related to insulin resistance and vitamin D deficiency in polycystic ovary syndrome. *European journal of endocrinology* **164**:741-749.

Wei S, Wang L-E, McHugh MK, Han Y, Xiong M, Amos CI, Spitz MR and Wei QW (2012) Genome-wide gene–environment interaction analysis for asbestos exposure in lung cancer susceptibility. *Carcinogenesis* **33**:1531-1537.

Weiskopf NG, Rusanov A and Weng C (2013) Sick patients have more data: the non-random completeness of electronic health records, in *AMIA Annual Symposium Proceedings* p 1472, American Medical Informatics Association.

Weiskopf NG and Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20**:144-151.

Werner RM and Bradlow ET (2006) Relationship between medicare's hospital compare performance measures and mortality rates. *JAMA* **296**:2694-2702.

Wilcox C, Feddes G, Willett-Brozick J, Hsu L-C, DeLoia J and Baysal B (2007) Coordinate up-regulation of TMEM97 and cholesterol biosynthesis genes in normal ovarian surface

epithelial cells treated with progesterone: implications for pathogenesis of ovarian cancer. *BMC Cancer* **7**:223.
Willemse P, Van der Sijde R and Sleijfer DT (1990) Combination chemotherapy and radiation for stage IV breast cancer during pregnancy. *Gynecologic oncology* **36**:281-284.
Willer CJ, Dyment DA, Sadovnick AD, Rothwell PM, Murray TJ and Ebers GC (2005) Timing of birth and risk of multiple sclerosis: population based study. *BMJ* **330**:120.
Witsch-Baumgartner M, Ciara E, Löffler J, Menzel H, Seedorf U, Burn J, Gillessen-Kaesbach G, Hoffmann G, Fitzky B and Mundy H (2001) Frequency gradients of DHCR7 mutations in patients with Smith-Lemli-Opitz syndrome in Europe: evidence for different origins of common mutations. *European journal of human genetics: EJHG* **9**:45-50.
Witsch-Baumgartner M, Sawyer H and Haas D (2013) Clinical utility gene card for: Smith-Lemli-Opitz Syndrome [lsqb]SLOS[rsqb]. *European journal of human genetics : EJHG* **21**.
Witsch-Baumgartner M, Clayton P, Clusellas N, Haas D, Kelley R, Krajewska-Walasek M, Lechner S, Rossi M, Zschocke J and Utermann G (2005) Identification of 14 novel mutations in DHCR7 causing the Smith-Lemli-Opitz syndrome and delineation of the DHCR7 mutational spectra in Spain and Italy. *Human mutation* **25**:412-412.
Wolff SM (1972) The ocular manifestations of congenital rubella. *Transactions of the American Ophthalmological Society* **70**:577.
Woodhouse P and Khaw K-T (2000) Seasonal variation of risk factors for cardiovascular disease and diet in older adults. *International journal of circumpolar health* **59**:204-209.
Xiang J, Nagaya T, Huang X-E, Kuriki K, Imaeda N, Tokudome Y, Sato J, Fujiwara N, Maki S and Tokudome S (2008) Sex and seasonal variations of plasma retinol, alpha-tocopherol, and carotenoid concentrations in Japanese dietitians. *Asian Pac J Cancer Prev* **9**:413-416.
Yamazaki JN and Schull WJ (1990) Perinatal loss and neurological abnormalities among children of the atomic bomb: Nagasaki and Hiroshima revisited, 1949 to 1989. *JAMA* **264**:605-609.
Yaris F, Yaris E, Kadioglu M, Ulku C, Kesim M and Kalyoncu NI (2004) Use of polypharmacotherapy in pregnancy: a prospective outcome in a case. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **28**:603-605.
Yu H and Patel SB (2005) Recent insights into the Smith–Lemli–Opitz syndrome. *Clinical Genetics* **68**:383-391.
Zamperini P, Gibelli B, Gilardi D, Tradati N and Chiesa F (2009) Pregnancy and thyroid cancer: ultrasound study of foetal thyroid. *Acta Otorhinolaryngologica Italica* **29**:339.
Zanobetti A and Schwartz J (2009) The effect of fine and coarse particulate air pollution on mortality: a national analysis. *Environ Health Perspect* **117**:898-903.
Zhang C, Qiu C, Hu FB, David RM, van Dam RM, Bralley A and Williams MA (2008) Maternal Plasma 25-Hydroxyvitamin D Concentrations and the Risk for Gestational Diabetes Mellitus. *PLoS ONE* **3**:e3753.
Zhou J, Ito K, Lall R, Lippmann M and Thurston G (2011) Time-series analysis of mortality effects of fine particulate matter components in Detroit and Seattle. *Environmental health perspectives* **119**:461.
Zoëga H, Valdimarsdóttir UA and Hernández-Díaz S (2012) Age, academic performance, and stimulant prescribing for ADHD: a nationwide cohort study. *Pediatrics* **130**:1012-1018.
Zou L and Porter TD Rapid suppression of 7-dehydrocholesterol reductase activity in keratinocytes by vitamin D. *The Journal of Steroid Biochemistry and Molecular Biology*.

Zuger A (2015) Hospital Ratings: A Guide for the Perplexed. *JAMA* **313**:1911-1912.